Gender Specific Impact of Replication and Recombination on Rodent Intron Evolution

Pink, Catherine

Award date:
2012

Awarding institution:
University of Bath

Gender Specific Impact of Replication and Recombination on Rodent Intron Evolution

Catherine Jennifer Pink

A thesis submitted for the degree of Doctor of Philosophy

University of Bath

Department of Biology and Biochemistry

February 2012

Contents

Acknowledgements

Collaborator contribution

All work presented in this thesis is my own, with the following exceptions:

The rat Y-linked sequence used in Chapters 2 and 3 was obtained and curated by Siva K. Swaminathan under the supervision of Andrew Ward, with sequencing performed by Ian Dunham and Jane Rogers at the Wellcome Trust Sanger Institute. Orthologous mouse Y-linked sequences were identified and curated by Siva K. Swaminathan and Nizar N. Batada.

The analysis of focal-flanking genes in Chapter 2 was performed by Laurence D. Hurst using data I had generated.

Chapters 2 and 3 were written by myself in collaboration with Laurence D. Hurst.

The work in Chapter 4 and the associated publication was conceived and primarily written by myself, with comments and suggestions from Laurence D. Hurst.

Curation of the replication timing data described in Appendix 1 was exclusively my responsibility. Analyses of these data were performed by Claudia C. Weber and Laurence D. Hurst, who also wrote the associated publication: WEBER, C. C., PINK, C. J. & HURST, L. D. (2012) Late-Replicating Domains Have Higher Divergence and Diversity in *Drosophila melanogaster*. *Molecular Biology and Evolution*. 29(2), 873–882.

Appreciation and dedication

This thesis is dedicated to my grandfather. I missed being able to discuss it with him.

Most of all I thank my family, who have supported and encouraged me throughout my university career. I thank my supervisor Laurence for pushing me as an undergraduate, for introducing me to bioinformatics and for giving me the opportunity to do this PhD. As Rosie, Toby, Elaine and Nicola venture on to bigger and better things, I shall miss working with them all greatly. The unsung people who keep the biology department running, particularly Amanda, Barbara, Emma, Teresa, Martin, Helen, Allun, Chris, Ulrika and Sandra, also deserve a huge thank you. My work was funded by a Capacity Building Studentship from the Medical Research Council.

Finally, I thank Chloe. I wish she could have completed the journey with me.

Abbreviations

A	Adenine
ANCOVA	Analysis of covariance
bp	Base pair
BrdU	The nucleotide analog 5-bromo-2-deoxyuridine
C	Cytosine
CI	Confidence interval
DNA	Deoxyribonucleic acid
DSB	Double strand break
G	Guanine
gBGC	GC biased gene conversion
GC	Guanine + cytosine content
GC*	Stationary GC content
Kb	Kilobase
K_i	Intronic substitution rate
K_S	Synonymous substitution rate
K_4	Substitution rate at four-fold degenerate sites
Mb	Megabase
MGI	Mouse Genome Informatics
NCBI	National Centre for Biotechnology Information
PAR	Pseudoautosomal region
PCR	Polymerase chain reaction
RGSPC	Rat Genome Sequencing Project Consortium
RI	Rearrangement index
RR	Recombination rate
RT	Replication time
SEM	Standard error of the mean
SNP	Single nucleotide polymorphism
T	Thymine
TEs	Transposable elements
UCSC	University of California Santa Cruz Genome Bioinformatics

Summary

Mutation rate variability has been widely observed across mammalian genomes but the underlying causes are not yet fully understood. This thesis attempts to explain this variation, as assayed by the substitution rate of putatively neutral sites, across rodent genomes at three scales: genic, inter-autosomal and between chromosome types. It was shown that the method commonly employed to estimate the extent of male-bias in the mutation rate is flawed, suggesting that inter-chromosomal variation in mutation rates is not solely due to differences in the number of replications they undergo in each germ-line. Two novel models were proposed that incorporated an additional recombination-associated parameter to explain why, contrary to the theory of male-driven evolution, the autosomes evolve faster than the Y-chromosome. As number of replications could not fully account for mutational variability at any scale, the impact of the time during S-phase when replication occurs was explored. Differential timing of replication was shown to explain both inter-genic and some inter-autosomal variation in intronic substitution rates, with later replicating sequences evolving faster. However, controlling for different replication times failed to account for why number of replications could not explain differences in chromosomal divergence. Further, GC rich sequences were found to evolve slowly because they tend to replicate early. Finally, late replicating genes were found to have high recombination rates in females but low recombination rates in males. These previously unidentified relationships could explain why, owing to sex-specific covariance with replication timing, the strength of covariance between recombination rate and divergence was underestimated in males and overestimated in females. It might also explain why female recombination rates, unlike those in males, do not covary with GC content.

Chapter 1. **Introduction**

Heritable mutations, those that arise in the germ-line, are essential to evolution, giving rise to the variation on which natural selection can act. Understanding where, when and why new mutations arise is therefore fundamental to many aspects of scientific enquiry: it informs our understanding of the underlying molecular mechanisms giving rise to new mutations; it enables accurate use of molecular clocks to determine phylogenetic relationships between species; it allows us to identify regions of the genome that might be more susceptible to new mutations and therefore to predict any disease effects that might result from this; it provides the comparison against which we identify the strength and direction of selection; and it can show us how the genome responds to such an influx of variation. How do these mutations arise and are some regions of the genome, particularly mammalian genomes, more susceptible to mutation than others?

1.1 Male driven evolution

Haldane (1947) was the first to propose that mutations might arise due to faulty copying of genes at cell division. He suggested that this might explain the discrepancy he observed in the mutation rate between males and females, given that spermatogonia continue to divide throughout an adult male's life, whereas in females, oocytes are almost formed by birth. This gives rise to a different number of cell divisions, and as we now know, DNA replications, between the two sexes. His proposed higher mutation rate in males than in females has since been termed 'male driven evolution'. Quantification of this hypothesis was made possible by Miyata and colleagues (1987b). They combined Kimura's (1968) proposal that the majority of mutations are selectively neutral and that the rate of molecular evolution must therefore be equal to the mutation rate, with a simple observation: that different types of chromosome differ in the frequencies that they spend in each germ-line - the Y chromosome being restricted to males; the X chromosome being twice as likely to be in a female than a male and the autosomes on average spending equal amounts of time in each germ-line. Given this, different chromosomal types would be exposed to the higher mutational input of the male germ-line and the lower mutational input

of the female germ-line with different frequencies and should therefore evolve at different rates. By comparing the per nucleotide rate of neutral evolution of any two types of chromosome, Miyata *et al*. (1987b) proposed a series of equations that enabled the ratio of male to female mutation rates, and thus the extent of male bias in the mutation rate, to be determined. Where no sex-bias was found, this ratio, termed α, would be equal to 1 whereas $\alpha > 1$ would be indicative of a male bias. In the first application of this model to a comparison between human and rodents, they found that male specific Y-linked sequence did indeed evolve faster than that of the autosomes, and in turn that autosomal sequence evolved faster than sequences located on the X chromosome (Miyata *et al*., 1987b, Miyata *et al*., 1987a).

Haldane's hypothesis was based on differences between the two sexes in the number of DNA replications that occur in the germ-line. As spermatogonia must be maintained for longer in species that reproduce later in life, this would therefore predict that species with a longer generation time should have a stronger male bias to the mutation rate. Using Miyata *et al*.'s (1987b) equations, subsequent work attempted to test this prediction by calculating the extent of male bias in the mutation rate in a number of species. Early work, based on comparisons of small amounts of often homologous sequences, found that estimates of α were consistent with what might be expected given the generation time (between birth and reproduction) with a stronger male bias in primates (Makova and Li, 2002, Shimmin *et al*., 1993) compared to rodents (Sandstedt and Tucker, 2005, Chang *et al*., 1994) or flies (Bauer and Aquadro, 1997).

The theory of male driven evolution has become one of the dominant explanations for the origin of new mutations (Ellegren, 2007, Crow, 1997), with new work in the field tending to focus on estimating α in non-model species (e.g. Ellegren and Fridolfsson, 2003, Berlin *et al*., 2006, Nakagome *et al*., 2008), relating α to life history characteristics (Sayres *et al*., 2011) or extending the hypothesis to include mutations types other than point mutations (Makova *et al*., 2004, Sundström *et al*., 2003). However, even as Miyata *et al*. (1987b) formulated their equations, variability in the mutation rate was being detected at different scales across genomes. Wolfe, Sharpe and Li (1989) found that synonymous substitution rates at four-fold

degenerate sites varied across rodent genes, an observation that has been confirmed by many studies (e.g. Bernardi *et al*., 1993, Wolfe and Sharp, 1993, Matassi *et al*., 1999). Later, Gaffney and Keightley (2005) showed that in rodent genomes, the scale of mutational variability was likely to be sub-chromosomal, probably around 1Mb. On a broader scale, different autosomes have been found to have significantly different rates of evolution in both primates and rodents (Malcom *et al*., 2003, Lercher *et al*., 2001). This mutational variability cannot be explained by differences in the number of DNA replications occurring in the male and female germ-lines, since across the autosomes this should be the same.

This autosomal variability raises an important concern: If the model can not explain the observed regionality of mutation rates, then this must call into question the accuracy of the model itself and suggest that its underlying hypothesis, that random errors during DNA replication are the dominant cause of mutational variability, may be incorrect.

Miyata *et al*.'s (1987b) model has also been called into question by conflicting estimates of α generated by numerous studies, where different chromosomal comparisons of different sequence types have been employed. In rodents, for example, α derived from X to autosomal comparisons of synonymous sites have ranged from 3.5 (Malcom *et al*., 2003) to ∞ (Wolfe and Sharp, 1993, McVean and Hurst, 1997) whereas X to Y comparisons of intronic sites have given estimates of α around 2 (Sandstedt and Tucker, 2005, Chang *et al*., 1994, Chang and Li, 1995). Y to autosomal comparisons have yielded estimates of $\alpha \approx 1$ (McVean and Hurst, 1997). In the first study to apply all three of Miyata *et al*.'s (1987) equations in a single analysis, Smith and Hurst (1999) failed to retrieve a single value of α. Depending on the type of sequence used to determine rates of evolution, estimates varied with $\alpha_{XA} = 8.96 - \infty$, $\alpha_{XY} = 1.79 - 2.99$ and $\alpha_{YA} = 0.71 - 1.95$.

That more new mutations are generated in males than in the female germ-line has further been called into question by medical evidence. For dominant autosomal disorders, when unaffected parents have an affected child, the ratio of paternal to maternal origins of the new mutation can be used to estimate α. Studies involving

multiple endocrine neoplasia (MEN), achondroplasia and Apert syndrome have shown an exclusively paternal origin for the disease causing mutation, giving α of ∞, consistent with the male driven evolution hypothesis (see Hurst and Ellegren, 1998 and references therein). However, for Hirschsprung disease, a disorder that involves the same gene as MEN, all three characterised mutations have been found to be maternal in origin (again see references within Hurst and Ellegren, 1998). Intriguingly, a recent study by Conrad *et al*. (2011) sequenced the complete genomes of each member of two families and found that the direction of a sex bias was dependent on the family, with 92% of de novo mutations arising in the paternal germ-line in one family, but 64% being of maternal origin in the other.

Molecular examinations of the male-derived mutations giving rise to these diseases has also cast doubt on the role of additional spermatogonial replications in giving rise to male driven evolution. Goriely *et al*. (2003) examined the most common disease causing mutation giving rise to Apert syndrome, those at nucleotide 755 in the FGFR2 gene, and showed that mutation events were in fact relative infrequent, but that the strong male bias associated with Apert syndrome could instead be attributed to a selective advantage to the spermatogonial cells in which the mutations arose. They suggested that this might be owing to a gain of function in the encoded protein that, in the cellular context of the testis, is advantageous, despite later being harmful to the developing embryo. The group later showed that mutations at the 755 nucleotide, particularly rare double substitutions at the same site, conferred a proliferative advantage that resulted in a clonal expansion of mutation carrying cells in the testis relative to neighbouring wild type cells. As mutation rates at adjacent nucleotides were found to be low, this suggested that their results could not be attributed to a mutational hotspot (Goriely *et al*, 2005). Consistent with these findings, Qin *et al*. (2007) used a novel approach to determine the stage of spermatogenesis at which mutations predominantly arise. By examining the spatial distribution across the testis of the C755G mutation in the FGFR2 gene, they also showed that the high incidence of paternal mutations at this site was not owing to a mutational hotspot. Instead, rather than just replacing themselves with one copy per mitotic cell division (the other cell going on to differentiate into a mature gamete), the adult self-renewing spermatogonial cells carrying the mutation occasionally

produced two SrAP cells. These studies would suggest that, certainly for Apert syndrome, evidence in favour of a male mutation bias may have been misinterpreted. It is worth speculating that a similar selective advantage, such as the genome containing a novel mutation being less likely to be diverted to the polar body, might explain the female bias found for Hirschsprung disease.

Another possible explanation for the apparent male bias to the mutation rate was suggested by McVean and Hurst (1997), who argued that selection against weakly deleterious recessive mutations on the X chromosome, when exposed in males, could potentially reduce the rate of X-linked evolution. Comparison of either the autosomal or Y-linked rate of evolution with the reduced X-linked rate would therefore give results suggestive of an elevated male mutation rate. Consistent with their hypothesis, McVean and Hurst found that the Y-linked synonymous substitution rate did not differ significantly from that of the autosomes. Note however, that this hypothesis does not necessarily suggest Miyata *et al*.'s (1987b) model is flawed, but rather that it has been incorrectly applied: the use of sites under selection to estimate substitution rates does not provide an accurate measure of the mutation rate.

Why else might the evidence not universally support the theory of male driven evolution? Fundamentally, the hypothesis is based on the concept that the majority of new heritable mutations arise as errors during DNA replication in the germ-line. However, a number of other covariates of mutation rates have been identified.

1.2 GC content

GC content, the proportion of G and C bases relative to A and T bases, is known to vary across mammalian genomes giving rise to the well documented isochore structure (Bernardi *et al*., 1985). This variation affects both coding and non-coding sites over scales ranging from kilobases to megabases. A number of explanations for this variation have been proposed. Variability in the efficiency of DNA repair or the mutability of different sequences are known, but these are not thought to operate over the scales required to give rise to the broad scale isochore structure, nor give

rise to patterns of isochore decline observed in mammals (Duret *et al*., 2002). Alternatively, natural selection owing to thermal stability (Bernardi, 2000) or amino acid composition (D'Onofrio *et al*., 1991) has been suggested, although these explanations have largely been ruled out as they are not supported by the origin of isochore structure relative to that of homeothermy (Belle *et al*., 2002) or why GC content should vary in non-coding sequences respectively (Eyre-Walker and Hurst, 2001). As will be discussed later in this chapter, the current favoured explanation is that of recombination-associated biased gene conversion.

That G and C are approximately twice as mutable than A and T was suggested some time ago (Gojobori *et al*., 1982, Bulmer, 1986). However, whether there is a relationship between GC content and mutation rates and if so, what the nature of this relationship might be, has been the subject of contentious debate that has yet to be resolved. When Wolfe *et al*. (1989) identified mutational variation at four-fold degenerate sites, they found that it covaried positively with GC content in regions of the genome with high GC content, but that in AT rich regions, a negative covariance was found. This decline in substitution rates with increasing GC content in AT rich regions was confirmed for K_S (Filipski, 1987, Ticher and Graur, 1989). However, later studies (e.g. Bernardi *et al*., 1993) failed to recover any relationship between GC_4 and K_S. The recovery of an inverted v- or u-shaped relationship (Gu and Li, 1994) prompted further debate.

In an attempt to explain these discrepant relationships, Hurst and Williams (2000) applied a number of methodologies to the issue in rodents. Using a preferred maximum likelihood analysis they found a positive covariance between GC_4 and K_S that was best explained by a shallow u-shaped quadratic, but no relationship when the analysis was restricted to K_4. When different statistical models were applied they either failed to recover any relationship or found that the u-shape became inverted as had previously been reported. The positive covariance between GC_4 and K_4 has subsequently been confirmed (Lercher and Hurst, 2002), though here the better fit of a quadratic model was not tested. A comprehensive analysis between human and mouse (Hardison *et al*., 2003) extended the issue outside of coding regions and again recovered the u-shaped relationship between GC content and substitution rates at

both four-fold degenerate sites and in ancestral repeats. More recently, Tyekucheva *et al*. (2008) studied ancestral repeats in primates and again recovered a pronounced u-shaped covariance between GC and neutral rates, showing that a quadratic fitted the relationship better than a linear regression.

Many of the studies that recovered the u-shaped relationship between GC content and rates of neutral evolution have considered the impact of highly mutable CpGs. Mutation rates at CpG sites are elevated 10-fold, driven mostly by an elevation in the transition rate, though transversion rates also increase (Hodgkinson and Eyre-Walker, 2011). This effect is largely due to methylation of CpGs in mammals, the methylated C being more susceptible to deamination to thymine. Hellmann *et al*. (2005) suggested that the curved nature of the relationship in GC rich regions reflects an underlying quadratic relationship between GC content and the probability of finding a CpG site, that is to say that as GC content increases, it becomes increasing likely that such sites will form the highly mutable CpG residues. In GC rich regions, any increase in mutation rate with increasing GC content might therefore be attributable to CpG sites. Consistent with this, removal of CpG sites resulted in the loss of a relationship between GC content and neutral rates in regions of high GC content, whereas the negative relationship remained where GC content was lower (Tyekucheva *et al*., 2008). Hurst and Williams (2000) also showed that that some of the positive part of the relationship could be controlled by an excess of CpG to TpG or GpC to GpA mutations in GC rich regions.

1.3 Replication time

In their 1989 paper, Wolfe *et al*. (1989) suggested the possibility that changes in the dinucleotide pool during S-phase might affect mutation, given that it was known that the eukaryotic genome does not replicate synchronously. Until 2008, this hypothesis had remained untested. The ability to test this hypothesis was made possible by the production of genome-wide replication timing maps of increasing resolution, particularly in mammalian genomes (MacAlpine and Bell, 2005, Woodfine *et al*., 2004). Several methods to do this exist, one of which involves the use of a culture of asynchronously cycling cells that is pulse labelled with 5-bromodeoxyuridine

(BrdU). These cells are then sorted by flow cytometry into S-phase fractions based on DNA content, after which the BrdU-labelled DNA is immunoprecipitated from each fraction and amplified. The early and late fractions are then differentially labelled and co-hybridized to a whole genome microarray. Relative replication times are calculated as a $\log_2$ ratio of early-to-late replication (e.g. Hiratani *et al*., 2008).

Despite conservation of origin binding proteins higher eukaryotes, unlike yeast, do not contain conserved autonomously replicating sequences where replication originates. However, during G1 phase in all eukaryotes, prereplication (preRC) complexes assemble at multiple origins of replication so that they are licensed - capable of initiating DNA replication. Not all of these origins are activated during S-phase and those that do are not activated synchronously. Timing of replication of a given sequence is therefore a function both of distance from an active origin and the time that the origin was activated. The order in which these origins are activated is not random. Instead, replication is organised into clearly defined regions of similarly timed replication. Origins firing synchronously tend to cluster, forming discrete replication foci and in turn these foci or 'replicons' may form larger domains of similarly timed replication that range from a few kilobases to several megabases in size (Hiratani *et al*., 2008).

These replicons and domains are relatively stable, with foci that initially replicated at a single time point continuing to do so in subsequent S-phases down a cell line (Jackson and Pombo, 1998). Both sharp and gradual transitions between replication domains have been reported (Woodfine *et al*., 2004, Farkash-Amar *et al*., 2008), but in both cases transitions probably lack origins and their boundaries are shared between unrelated cell lines (Woodfine *et al*., 2004, Hiratani *et al*., 2008). Replication timing in mammals has been shown to correlate between different cell lines (Woodfine *et al*., 2004), although recent higher resolution studies have shown that as much as 20% of the genome's replication timing is subject to change upon cell differentiation, upon which new domains were again conserved between unrelated cell lines of a similar cell type, suggesting that replication timing is characteristic of a specific cell type (Hiratani *et al*., 2004, Hiratani *et al*., 2008). Indeed, the strength of this relationship is so strong that once mapped, replication

timing appears to be a unique identifier that can be used to predict cell type (Ryba *et al*., 2011). Using similar cell types, homologous genes in human and mouse have been found to replicate at similar times, despite sequence divergence, rearrangements, and even when other features such as GC content and transcription differed (Yaffe *et al*., 2010, Farkash-Amar *et al*., 2008). Together, these findings suggest that, within mammals at least, replication timing is a relatively stable genomic feature that is conserved during evolution.

Since their discovery, replication domains have been found to have a number of important characteristics, though whether these are determinants of or stem from differences in replication timing remains a subject of investigation. Early replication timing has been associated with high GC content, high gene density, gene expression, low LINE repeat sequence density and, in humans, high Alu repeat density (Woodfine *et al*., 2004). However, as these features also tend to covary with each other, the extent to which replication time is involved is not yet certain.

GC content has long been associated with differences in replication timing (Schmegner *et al*., 2007, Costantini and Bernardi, 2008) and remains its strongest covariate (Woodfine *et al*., 2004). This relationship is not, however, static, with the aforementioned changes during cell differentiation tending to bring replication timing into alignment with isochore structure, such that previously early replicating AT rich regions become late replicating and *vice versa* (Hiratani *et al*., 2008).

The nature of the relationship between timing of replication and gene expression has been the subject of much research. Links between the two appear to be restricted to higher eukaryotes, none being found in yeast (Raghuraman *et al*., 2001) but being found in *Drosophila* (Schübeler *et al*., 2002, MacAlpine and Bell, 2005) and mammals (Woodfine *et al*., 2004, Hiratani *et al*., 2008, Farkash-Amar and Simon, 2009). In multicellular species, housekeeping genes that are expressed ubiquitously are known to replicate early in mammals whereas those that are tissue specific tend to replicate late except where expressed (Holmquist, 1987, Farkash-Amar *et al*., 2008, Selig *et al*., 1992). Further, genes that alter their replication timing from early to late S-phase tend to be down-regulated, whereas those becoming early replicating

have a tendency to increase their expression levels (Hiratani *et al*., 2008). There are, however, exceptions to this rule that suggest a more complex relationship, with repressed genes replicating early and expressed genes replicating late (Gilbert, 2002, Farkash-Amar *et al*., 2008). This is not restricted to mammals - in *Drosophila*, 30% of early replicating genes are inactive and 30% of active genes are late replicating (Schwaiger *et al*., 2009).

In humans, the probability of expression, that is, whether or not a gene is expressed at all, is a stronger covariate of replication timing than absolute level of expression (Woodfine *et al*., 2004). Consistent with this, replication timing in mouse has been found to be related to RNA polymerase promoter occupancy, with genes that are primed for but not being actively transcribed replicating early (Farkash-Amar *et al*., 2008). Further, genes involved in stress response and apoptosis, which are likely to be inactive but require rapid transcription, replicate early (Farkash-Amar *et al*., 2008). These results tend to suggest that genes primed for expression but not undergoing active transcription replicate early. Combined with comparisons of pluripotent to differentiated cells, this model suggests that much of the genome initially replicates early, with many genes primed for possible expression. Then, as differentiation occurs, expression of subsets of genes that are no longer required is shut down and they move to later replication times in the terminally differentiated cells (Hiratani *et al*., 2008). Notably, the relationship between expression and replication timing does not extend to a subset of genes with strong promoters, enabling late replicating genes to be upregulated.

The causal relationship between expression and replication timing was initially thought to be due to transcriptionally silent genes being inaccessible to replication factors, thus delaying replication. However, evidence now favours a model whereby replication timing determines chromatin structure, this being formed at the replication fork, and that chromatin structure then determines transcriptional potential (Farkash-Amar *et al*., 2008). Changes in replication timing might therefore facilitate changes in chromatin structure and subsequently expression potential. This would be supported by the observation that chromatin permissive for expression replicates early, but heterochromatin tends to be one of the last types of DNA to

replicate (Gilbert, 2002). Further, while the human β-globin gene is usually inactive and late replicating, but early replicating when it is expressed, experimental silencing of the gene whilst maintaining the open chromatin state results in early replication (see Schwaiger *et al*., 2009). Again though, there are exceptions, with bivalent genes, those exhibiting co-occupation of active and inactive histone modifications, always replicating early and their resolution not being linked to changes in replication timing.

Temporal variation in replication has also been linked to spatial variation, with early replication tending to occur in the centre of the nucleus whereas DNA located at the nuclear periphery replicates later (Dimitrova and Gilbert, 1999). Again, this feature of replication appears to be conserved, with discrete foci that replicate together continuing to occupy discrete chromosomal domains during interphase of subsequent cell generations (Jackson and Pombo, 1998). Again too, changes in replication time during differentiation have been associated with spatial changes, with genes moving either towards or away from the nuclear periphery (Hiratani *et al*., 2008).

When Wolfe *et al*. (1989) initially suggested that replication timing might be a determinant of mutation rates, they speculated that this might be due to changes in the composition of the dinucleotide pool during S-phase, although they incorrectly supposed an increase in GC content in later replicating DNA. Intracellular concentrations of deoxyribonucleoside triphosphates (dNTPs) are tightly controlled and regulated. At the G1/S-phase border dNTP pools start to increase and continue to expand during S-phase. When S-phase is completed, production of dNTPs is shut off and a residual pool left for DNA repair outside of S-phase (Mathews, 2006). In mammals, this mirrors changes in the speed of replication, which starts slowly, increases to a linear rate a third of the way through S-phase, during which most of genome is replicated, before possibly starting to slow (Woodfine *et al*., 2004, Farkash-Amar *et al*., 2008).

Critical to replication, changes in dNTP availability have been shown to influence the speed of fork progression in early, but not late, S-phase (Malínsky *et al*., 2001) and origin activation in mammals (Anglana *et al*., 2003). Reduction in dNTP

concentration below critical levels has been linked with replication arrest (Koç *et al*., 2004) and, when dNTP pools are depleted, initiation of the S-phase check point (Kumar *et al*., 2010). In nearly all organisms studied, dNTP pools are asymmetric with an excess of dATP and dTTP and an under-representation of dGTP, the latter comprising only 5-10% of the pool (Mathews, 2006). There is also evidence of spatial variation: while dCTP is compartmentalised and is subject to the greatest nuclear increases in concentration during S-phase, dTTP exists in a single equilibrated pool (Leeds *et al*., 1985, but see Kumar *et al*., 2011, Xu *et al*., 1995).

It is generally assumed that dNTP concentrations have evolved to maximise replication fidelity. Consistent with this, perturbation of relative concentrations to either equimolar conditions or extreme imbalances has been shown to be mutagenic (Martomo and Mathews, 2002, Kumar *et al*., 2011, Kumar *et al*., 2010). Similarly, a balanced accumulation of dNTPs is also mutagenic, possibly due to a reduction in proof reading due to saturation of the DNA polymerase by dNTPs (Martomo and Mathews, 2002). Mechanisms by which an excess of dNTPs are thought to lead to mutations include competition between dNTPs resulting in mismatches, with the nature of the mutation being dependent on the direction of the imbalance; frameshifts due to realignments of sequence after excess dNTPs form a correct pairing at a slipped site; and excess dNTPs driving chain extension past mismatched sites before the polymerase is able to detect and correct the error (Mathews, 2006). However, while imbalances in dNTP pools are known to occur in cancerous cells, it is not yet known whether imbalances and their associated mutagenic effects are a feature of normally functioning cells.

1.4 Recombination

Recombination rates are known not be uniform across the genome. In addition to a broad scale elevation in recombination rate towards telomeres and suppression at centromeres, finer scale variation also exists. Most recombination events are concentrated in short regions, termed ‘recombination hotspots’. In mammals, recombination hotspots are estimated to be 1-2 Kb in size and tend to occur at 50-100 Kb intervals (Myers *et al*., 2006). Numerous recombination hotspots have now

been well characterised in a number of species (Jeffreys and Neumann, 2009, Wu *et al*., 2010). In addition to variation in recombination rates within a given genome, variation also exists between species (Ptak *et al*., 2005, Jensen-Seaman *et al*., 2004), between genders (Paigen *et al*., 2008, Kong *et al*., 2002, Jensen-Seaman *et al*., 2004) and even between individuals of the same species (Dumont *et al*., 2011, Coop *et al*., 2008).

Understanding the causes of this recombination rate variation demands knowledge of the underlying mechanisms that determine where recombination occurs. Substantial progress has been made in this field over recent years, with the discovery both of a degenerate 13 bp consensus sequence enriched in over 40% of human recombination hotspots (Myers *et al*., 2008), and of a chromatin modifying protein, PRDM9, that is predicted to bind to these motifs (Baudat *et al*., 2010, Myers *et al*., 2010, Parvanov *et al*., 2008). This protein results in the trimethylation of lysine 4 in histone H3, which in yeast and mice has been found to define the sites at which double strand breaks are formed during normal meiosis (Borde *et al*., 2009, Buard *et al*., 2009) – these double strand breaks being required to initiate chiasma. Studies have shown that the mouse ortholog of PRDM9, a zinc finger protein, is not only expressed during meiotic prophase but that knock out strains fail to repair double strand breaks (DSB) and are infertile (Myers *et al*., 2010, Parvanov *et al*., 2010), strongly supporting its role in recombination.

Further evidence in support of this model comes from its ability to explain the observed variation in recombination rate. PRDM9 has been found to evolve extremely rapidly between human and chimpanzee, and across a number of mammals, explaining why similar consensus motifs do not induce recombination in different species (McVean and Myers, 2010, Oliver *et al*., 2009). Further, variation in the number of PRDM9 zinc finger domains between different mouse strains, together with amino acid substitutions at key DNA binding sites, would explain variation between their genetic maps (Parvanov *et al*., 2010). Variation in PRDM9 has also been observed within humans, with a variant found in humans of European ancestry not being associated with the common motif (Baudat *et al*., 2010). Similarly, the probability of a crossover in recombination hotspots active in people

of West African ancestry, but inactive in Europeans, appears to be controlled by the PRDM9 allelic variant and enrichment for the associated binding motif within the hotspot (Hinch *et al*., 2011). This rapid evolution of PRDM9 might reflect a need to adapt to disruptions of the consensus motif via biased gene conversion, leading to hotspot loss (Myers *et al*., 2010).

In contrast to variation between species and individuals, the variation in the distribution of recombination hotspots between genders has not yet been attributed directly to diversity in the PRDM9 protein. It has long been known that the female genetic map is longer than that of males – 1.7 times as long in humans and 1.3 times as long in mice (Kong *et al*., 2002, Shifman *et al*., 2006, Broman *et al*., 1998). This has been attributed to the formation of a longer synaptonemal complex in females (Tease and Hulten, 2004). There is also broad scale variation along chromosomes, with recombination events tending to be more strongly localised in males compared to a relatively even distribution in females. Male recombination rates also tend to increase towards the telomeres whereas recombination rates are higher around the centromeres in females (Paigen *et al*., 2008, Kong *et al*., 2002). Gender differences also exist in the fine scale location of recombination hotspots (Kong *et al*., 2010). However, a recent update of the genetic map of mouse has shown that gender differences, while still visible, were not as extreme as previously thought (Cox *et al*., 2009). One possible mechanistic explanation for these differences is that females have a more compact chromatin structure during the pachytene stage of meiosis and that this then results in a shorter genomic interference distance (Petkov *et al*., 2007).

Alternatively, a selective explanation has been proposed for the evolution of heterochiasmy, the gender differences in overall recombination rate. This suggests that recombination might be avoided in order to prevent breaking epistatic interactions between linked loci that selection has brought together. Such genes are particularly likely to be expressed during the haploid stage of either germ-line. As the haploid phase of the female germ-line is virtually non-existent – the final meiotic division taking place post fertilisation – this selective pressure would be stronger in the male germ-line, resulting in the lower recombination rate generally observed in males compared to females (Lenormand, 2003, Lenormand and Dutheil, 2005).

However, an argument could also be made for increased recombination in such regions. This would bring together new combinations of alleles that may be beneficial when linked in a highly competitive environment, such as sperm competition in males. In contrast, this hypothesis would predict an elevated recombination rate in males, or at least at genomic locations containing genes involved in sperm production, motility and fertilization (Lenormand, 2003, Lenormand and Dutheil, 2005). Different epistatic interactions between males and females would also explain why imprinted genes have elevated recombination rates (Lercher and Hurst, 2003).

Magni (1963) was the first to propose that meiotic recombination might be mutagenic. Since then, a number of studies have found a relationship between neutral rates of evolution and meiotic recombination. Lercher and Hurst (2002) suggested that recombination is mutagenic, based on observed elevated SNP diversity in regions of the human genome with high recombination rates. Observations of this type can also be attributed to hitchhiking or background selection, both of which require proximity to an allele under selection. However, there is also a covariance of SNP density and recombination in non-coding sequence, thought not to be under such selective effects (Lercher and Hurst, 2002). This would account for why Nachman (2001) failed to recover a similar relationship, since this analysis was restricted to sequence in close proximity to exons. A positive correlation between recombination and K_4 was also observed (Lercher and Hurst, 2002). This latter relationship was also detected at a 1Mb scale by Hardison *et al*. (2003), who also showed that recombination positively covaried with evolutionary rates in transposable elements. More recently, Tyekucheva *et al*. (2008) found that the human neutral rate in ancestral repeats covaries with male-, but not female-specific recombination rates.

Strong evidence for a recombinational-mutagenic effect comes from the pseudoautosomal region that, during the obligatory pairing of the X and Y chromosomes during male meiosis, is the only part of the Y chromosome to undergo inter-chromosomal recombination. As such, the pseudoautosomal region experiences rates of recombination that are 7 to 10 fold higher during male meiosis than when

recombination can occur along the length of the X chromosome during female meiosis (Soriano *et al*., 1987, Perry and Ashworth, 1999, Lien *et al*., 2000). This region has been shown to evolve extremely rapidly. Genes that are pseudoautosomal in humans but autosomal in rodents have been found to evolve faster (Ellison *et al*., 1996). Similarly, intronic sequences in the pseudoautosomal region of the human p-arm were found to evolve significantly faster than the genomic average. In contrast, intronic divergence in the q-linked pseudoautosomal region, which does not experience the same elevated recombination rate, did not differ significantly from the rest of the genome (Filatov and Gerrard, 2003).

A particular example is that of the *Fxy* gene. This gene moved to straddle the boundary of the pseudoautosomal region in laboratory mice so that 3 of the 10 exons are now located in the pseudoautosomal region, the rest remaining X-specific. Comparison with rat and other mice species, where the entire gene remains X-specific, has shown a huge acceleration in synonymous substitution rates that are 170-fold higher in the highly recombining pseudoautosomal exons compared to those which are only exposed to lower rates of recombination in females (Perry and Ashworth, 1999). Evidence has not however been universal. Like the mouse *Fxy* gene, the human and great ape XG gene straddles the p-pseudoautosomal region boundary. A study by Yi *et al*. (2004) showed that despite an elevated recombination rate, the rate of evolution of pseudoautosomal XG introns did not differ from those that are X-specific. Galtier (2004) proposed that a shift in the position of the boundary, such that in the ancestor of the great apes the entire XG gene was pseudoautosomal, might explain this discrepancy. This would also explain why Filatov (2004) was able to detect a gradient of increasing intronic substitution rates extending from the human pseudoautosomal boundary to towards the telomere.

One potential mechanism that has been proposed to explain this apparent mutational effect of recombination is that double strand breaks are mutagenic (Lercher and Hurst, 2002). After double strand breaks are initiated, DNA is degraded and then re-synthesised around the break (Figure 1.1). Unlike DNA replication, during DSB repair DNA synthesis is performed by low fidelity DNA polymerases (Rattray and Strathern, 2003, Strathern *et al*., 1995). In *Drosophila* however, not all DSBs result

in cross-overs, instead being resolved by alternative mechanisms such as synthesis-dependent strand annealing (Figure 1.1) or other methods of resolving Holliday junctions. It has been shown that in *Drosophila,* where crossing over is restricted extensive gene conversion tends to occur instead (Andolfatto and Wall, 2003, Langley *et al*., 2000). Given this Kulathinal *et al*. (2008) proposed that DSB repair, rather than crossing-over, is mutagenic, hence divergence in *Drosophila* often failing to correspond with genetic maps (Begun and Aquadro, 1992), these being measures of cross-over.

Nearly all studies examining the relationship between recombination and rates of evolution have found a positive relationship between recombination and GC content (e.g. Kong *et al*., 2002, Williams and Hurst, 2000, Eyre-Walker, 1993). Further, that recombination is driving GC content, rather than *vice versa*, comes from observations that recombination rate covarys more strongly with the GC content to which a sequence is evolving, rather than its current GC content (Meunier and Duret, 2004, Duret and Arndt, 2008). This has been attributed to the action of GC biased gene conversion (gBGC), which in turn is now the favoured explanation for the evolution of isochores (Romiguier *et al*., 2010, Duret and Galtier, 2009).

Gene conversion is the non-reciprocal copying of one homologous DNA sequence onto another and forms part of the meiotic recombination process (Figure 1.1). When the two parental chromosomes are heterozygous, mismatches form within the heteroduplexed gene conversion tract. Repair of these mismatches either reverts the mismatch to the original state or renders the chromosomes identical at the formally heterozygous site resulting in non-Mendelian segregation. In the case of the latter, if there is an equal probability of repair in either direction, on average the entire gamete pool remains unbiased. However, biases in the repair of T:G mismatches result in the formation of a higher proportion of G:C than A:T gametes. This neutral process generates a form of meiotic drive that gives GC alleles a transmission advantage over AT alleles (Figure 1.1). The main evidence for the mechanistic basis of this bias comes from a small number of studies that examined repair of G:T, A:C, C:T, A:G mismatches introduced into the SV40 viral genome in simian cells. The highest rate of repair was for G:T mismatches which tended to be corrected to G:C

(Brown and Jiricny, 1988). It is thought that this biased repair process might have evolved to correct G:T mismatches arising from cytosine deamination (Brown and Jiricny, 1987).

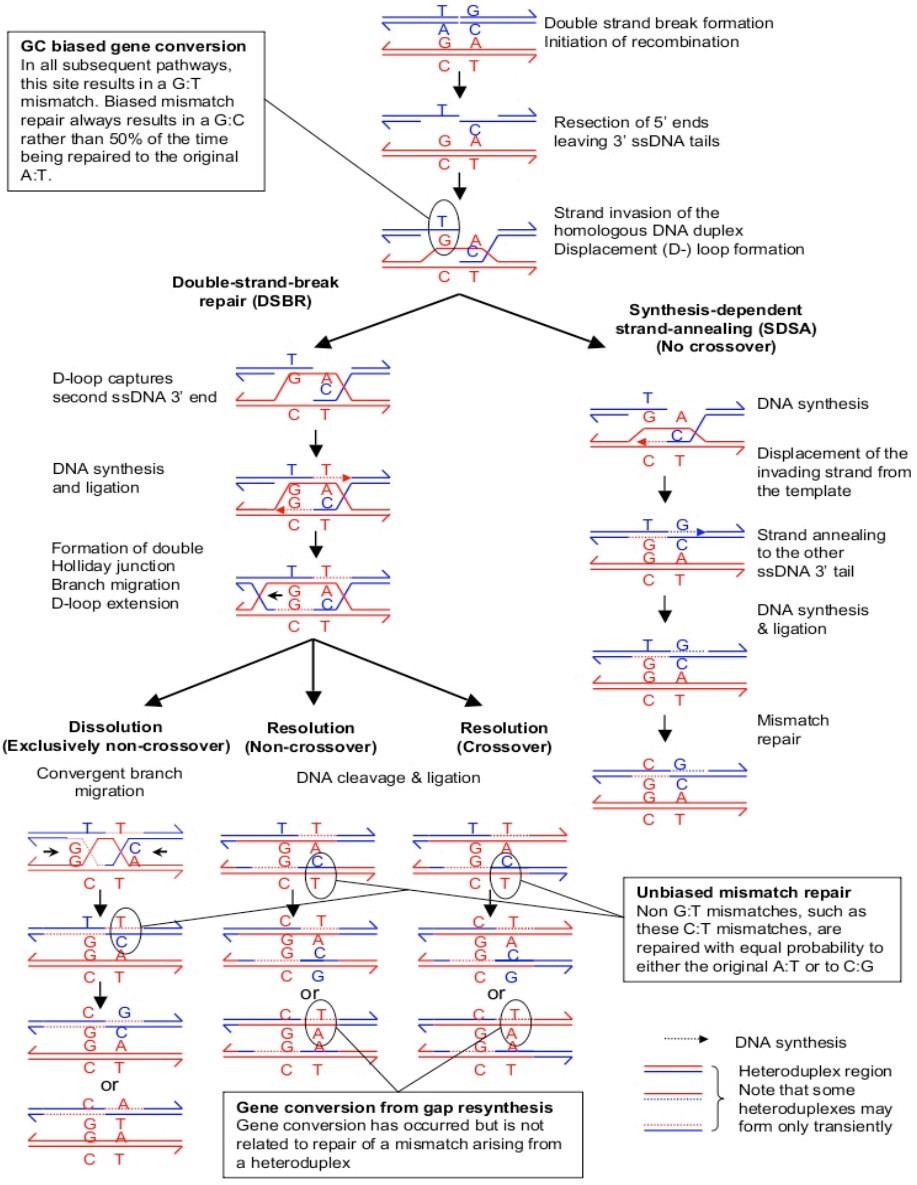


Figure 1.1: Alternative methods of repairing double strand breaks (DSBs) lead to GC biased gene conversion. Following resection, strand invasion is then resolved either by the synthesis dependent strand-annealing pathway or one of the double strand break repair pathways. Heteroduplexes form in all pathways, although some of these might only exist transiently, and where sites are heterozygous, mismatches occur in these regions. For most mismatches, such as the C:T mismatch shown here, mismatch repair either restores the original site or gene conversion to the other allele occurs. However, in the case of the G:T mismatch, repair is biased towards a G:C over A:T. This results in a transmission distortion in favour of GC alleles. Note that this diagram does not show complex gene conversion tracts. Information presented obtained from Chen *et al*. (2007), Duret and Galtier (2009) and Hurles (2001).

Returning again to the pseudoautosomal region, gBGC can account for why the pseudoautosomal portion of the domestic mouse *Fxy* gene has, in addition to the elevated substitution rate already described, a higher GC content than orthologous X-specific sequence (Perry and Ashworth, 1999). Again though, as for substitution rates, Yi *et al*. (2004) found that GC content did not differ between the pseudoautomal and X-specific parts of the XG gene, though notably both were elevated above the overall X-specific level. According to Galtier, this provides evidence in favour of recombination effect via gBGC. Movement into such a region would expose a sequence to the strong influence of a gBGC fixation effect, as observed in the mouse *Fxy* gene. In contrast, neutral mutation effects are much weaker, hence after removal of a sequence from strong recombinational environment, GC rich sequences are slower to return to GC content predicted by neutral effects alone (Galtier, 2004).

Could then the relationship between recombination and neutral rates of evolution be due to gBGC? Certainly, before a sequence reaches its stationary GC content, gBGC should result in an increased substitution rate. The rate at which recombination rates evolve might mean that this is a continuing process across the entire genome. Consistent with this, clusters of A/T (weak) to G/C (strong) human substitutions, but not SNPs, have been found to be common close to telomeres and covary more strongly with male than female recombination rates (Dreszer *et al*., 2007, Berglund *et al*., 2009). One explanation for why the latter might be is that recombination in males tends to be more clustered and evolve faster than in females. Further, the impact of gBGC might be so strong that it can overcome natural selection, promoting the fixation of deleterious AT→GC mutations or preventing advantageous GC→AT mutations from reaching fixation (Galtier *et al*., 2009, Berglund *et al*., 2009). For example, the human and short-tailed mouse versions of the *Fxy* protein differ by only six amino acids. However, since divergence from its rodent relative, the laboratory mouse *Fxy* protein has accumulated 28 amino acid changes, all of which are within the pseudoautosomal portion of the gene and all caused by AT→GC substitutions. Given the conservation across other species, these changes are unlikely to be adaptive (Galtier and Duret, 2007). However, evidence

from the pseudoautosomal region is not unanimous in support of a gBGC-associated substitution effect. Filatov (2004) suggested that the relationship between substitution rate and recombination in the pseudoautosomal region could not be due to gBGC alone. His examination of only A↔T or G↔C substitutions, those that could not be attributed gBGC, still resulted in a substitution rate gradient extending into the pseudoautosomal region.

1.5 Other sources of mutational variation

A number of other potential sources of mutational variation have been examined that will briefly be described here. As has been mentioned, features of replication timing and recombination are strongly associated with chromatin structure. This too has been examined as a possible covariate of mutation rates. Prendergast *et al*. (2007) found that in humans, intergenic and intronic divergence was highest in closed chromatin. They speculate that this might reflect higher rates of DNA damage, impaired DNA lesion detection or reduced DNA repair. However, Chen *et al*. (2010) later showed that the effect was in fact attributed to the later replication timing of heterochromatic regions. Whether chromatin structure is indeed a covariate, let alone casual, of mutation rate variability therefore remains unclear.

Mutational variability is also known to covary with indel occurrence. This is usually attributed to indirect relationships between either mutability of the sequence or strength of selection against all types of mutation (Hardison *et al*., 2003). However, that elevated substitution rates are particular to the orthologous sequence containing the indel, and that the scale of the effect, is proportional to indel size is suggestive of a causal relationship (Tian *et al*., 2008). It has been suggested that heterozygosity for the indel is mutagenic, disrupting pairing during meiosis, although notably the effect has not yet been associated with recombination hotspots in primates or yeast. Whatever the mechanism, the trend appears to be universal across eukaryotes, having been observed in primates, rodents, flies, rice and yeast (Tian *et al*., 2008).

Finally, gene expression has also been suggested as a possible source of novel mutations. Some evidence for this is indirect, with an observed clustering of genes

with both similar rates of evolution and similar levels of expression. For example, in both rodents and primates, introns of the same gene and linked genes have been found to evolve at more similar rates than expected by chance (Lercher *et al*., 2004). Evidence for a link between transcription and mutation comes also from strand asymmetry in base composition, where there is an excess of G over C and T over A on the coding strand of broadly expressed genes (Green *et al*., 2003, Majewski, 2003). That this coding strand excess of G and T is significantly higher in transcribed than in flanking untranscribed sequence (Green *et al*., 2003, Mugal *et al*., 2009) and shows clear transitions between equal and skewed base composition at the 5' end of transcribed DNA (Touchon *et al*., 2003, Touchon *et al*., 2004, Polak and Arndt, 2008) is strongly suggestive of a transcription associated effect. The cause of this strand asymmetry is thought result from the coding strand being transiently exposed in a single stranded state during transcription and as such, more susceptible to mutations such as depurination or deamination. Transcription-coupled repair then acts on the non-coding strand to repair the resulting mismatch.

1.6 Thesis aims

This thesis focuses on genome-wide mutational variability in the murid rodents mouse (*Mus musculus*) and rat (*Rattus norvegicus*). Chapter 2 aims to re-examine the primary explanation for the variation in rates of evolution between different types of chromosomes - the theory of male driven evolution. It expands on a previous study, making use of newly available Y-linked sequence, to explicitly test Miyata *et al*.'s (1987b) model. The chapter shows that Miyata *et al*.'s (1987b) model was unable to explain the data as, contrary to the prediction of the male driven evolution, the autosomal rate of intronic and exonic evolution exceeded that of the Y chromosome. Given previously observed relationships between recombination rate and divergence, the chapter also proposes a novel model that incorporates a recombination-associated substitution rate parameter.

An assumption of Miyata *et al*.'s (1987b) model is that, per replication, novel mutations are evenly distributed across the genome. Based on the discussed variability in replication during S-phase, Chapter 3 tests this assumption and shows

that intronic sequences that replicate at different times evolve at different rates. However, it is also shown that controlling for timing of replication is not able to account for the discrepancy in Miyata *et al*.'s (1987b) model shown in Chapter 2. It is also suggested that GC rich sequences might have low rates of evolution because they replicate early, rather than early replicating sequences evolving slowly because they are GC rich.

As discussed, early replication and high rates of recombination have both been associated with high GC content. Timing of replication and recombination rate might therefore be expected to covary. Chapter 4 tests this hypothesis and explores the relative impact that replication timing and sex-specific recombination rates have on both intronic divergence and GC content. The analyses show that late replicating domains tend to have high recombination rates in females but low recombination rates in males. As these trends are antagonistic, the relationship between recombination rate and divergence has been moderately underestimated for male recombination and slightly overestimated for female recombination, owing to covariance with replication timing. It also explains why male recombination is strongly correlated with GC content but female recombination is not, with GC promotion by biased gene conversion during female recombination being countered by the antagonistic effect of later replicating sequence tending to increase AT content.

1.7 References

ANDOLFATTO, P. & WALL, J. D. (2003) Linkage disequilibrium patterns across a recombination gradient in African *Drosophila melanogaster*. *Genetics,* 165**,** 1289-1305.

ANGLANA, M., APIOU, F., BENSIMON, A. & DEBATISSE, M. (2003) Dynamics of DNA replication in mammalian somatic cells: nucleotide pool modulates origin choice and interorigin spacing. *Cell,* 114**,** 385-394.

BAUDAT, F., BUARD, J., GREY, C., FLEDEL-ALON, A., OBER, C., PRZEWORSKI, M., COOP, G. & DE MASSY, B. (2010) PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science,* 327**,** 836-840.

BAUER, V. L. & AQUADRO, C. F. (1997) Rates of DNA sequence evolution are not sex-biased in *Drosophila melanogaster* and *D. simulans*. *Mol Biol Evol,* 14**,** 1252-1257.

BEGUN, D. J. & AQUADRO, C. F. (1992) Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. *Nature,* 356**,** 519-520.

BERGLUND, J., POLLARD, K. S. & WEBSTER, M. T. (2009) Hotspots of biased nucleotide substitutions in human genes. *PLoS Biol,* 7**,** e1000026.

BELLE, E. M. S., SMITH, N. & EYRE-WALKER, A. (2002) Analysis of the Phylogenetic Distribution of Isochores in vertebrates and a Test of the Thermal Stability Hypothesis. *Journal of Molecular Evolution,* 55**,** 356-363.

BERLIN, S., BRANDSTRÖM, M., BACKSTRÖM, N., AXELSSON, E., SMITH, N. G. C. & ELLEGREN, H. (2006) Substitution rate heterogeneity and the male mutation bias. *Journal of Molecular Evolution,* 62**,** 226-233.

BERNARDI, G. (2000) Isochores and the Evolutionary Genomics of Vertebrates. *Gene,* 241**,** 3-17.

BERNARDI, G., MOUCHIROUD, D. & GAUTIER, C. (1993) Silent Substitutions in Mammalian Genomes and Their Evolutionary Implications. *Journal of Molecular Evolution,* 37**,** 583-589.

BERNARDI, G., OLOFSSON, B., FILIPSKI, J., ZERIAL, M., SALINAS, J., CUNY, G., MEUNIER-ROTIVAL, M. & RODIER, F. (1985) The mosaic genome of warm-blooded vertebrates. *Science,* 228**,** 953-958.

BORDE, V., ROBINE, N., LIN, W., BONFILS, S., GÉLI, V. & NICOLAS, A. (2009) Histone H3 lysine 4 trimethylation marks meiotic recombination initiation sites. *The EMBO Journal,* 28**,** 99-111.

BROMAN, K. W., MURRAY, J. C., SHEFFIELD, V. C., WHITE, R. L. & WEBER, J. L. (1998) Comprehensive human genetic maps: individual and sex-specific variation in recombination. *Am J Hum Genet,* 63**,** 861-869.

BROWN, T. C. & JIRICNY, J. (1987) A specific mismatch repair event protects mammalian cells from loss of 5-methylcytosine. *Cell,* 50**,** 945-950.

BROWN, T. C. & JIRICNY, J. (1988) Different base/base mispairs are corrected with different efficiencies and specificities in monkey kidney cells. *Cell,* 54**,** 705-711.

BUARD, J., BARTHÈS, P., GREY, C. & DE MASSY, B. (2009) Distinct histone modifications define initiation and repair of meiotic recombination in the mouse. *The EMBO Journal,* 28**,** 2616-2624.

BULMER, M. (1986) Neighboring base effects on substitution rates in pseudogenes. *Mol Biol Evol,* 3**,** 322-329.

CHANG, B. H., SHIMMIN, L. C., SHYUE, S. K., HEWETT-EMMETT, D. & LI, W. H. (1994) Weak male-driven molecular evolution in rodents. *Proc Natl Acad Sci USA,* 91**,** 827-831.

CHANG, B. H. J. & LI, W. (1995) Estimating the Intensity of Male-Driven Evolution in Rodents by using X-linked and Y-linked UBE-1 Genes and Pseudogenes. *Journal of Molecular Evolution,* 40**,** 70-77.

CHEN, C.-L., RAPPAILLES, A., DUQUENNE, L., HUVET, M., GUILBAUD, G., FARINELLI, L., AUDIT, B., D'AUBENTON-CARAFA, Y., ARNEODO, A., HYRIEN, O. & THERMES, C. (2010) Impact of replication timing on non-CpG and CpG substitution rates in mammalian genomes. *Genome Research,* 20**,** 447-457.

CHEN, J.-M., COOPER, D. N., CHUZHANOVA, N., FÉREC, C. & PATRINOS, G. P. (2007) Gene conversion: mechanisms, evolution and human disease. *Nat Rev Genet,* 8**,** 762-775.

CONRAD, D. F., KEEBLER, J. E. M., DEPRISTO, M. A., LINDSAY, S. J., ZHANG, Y., CASALS, F., IDAGHDOUR, Y., HARTL, C. L., TORROJA, C., GARIMELLA, K. V., ZILVERSMIT, M., CARTWRIGHT, R., ROULEAU, G. A., DALY, M., STONE, E. A., HURLES, M. E., AWADALLA, P. & PROJECT, G. (2011) Variation in genome-wide mutation rates within and between human families. *Nat Genet,* 43**,** 712-714.

COOP, G., WEN, X., OBER, C., PRITCHARD, J. K. & PRZEWORSKI, M. (2008) High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans. *Science,* 319**,** 1395-1398.

COSTANTINI, M. & BERNARDI, G. (2008) Replication timing, chromosomal bands, and isochores. *Proc Natl Acad Sci USA,* 105**,** 3433-3437.

COX, A., ACKERT-BICKNELL, C. L., DUMONT, B. L., DING, Y., BELL, J. T., BROCKMANN, G. A., WERGEDAL, J. E., BULT, C., PAIGEN, B., FLINT, J., TSAIH, S.-W., CHURCHILL, G. A. & BROMAN, K. W. (2009) A new standard genetic map for the laboratory mouse. *Genetics,* 182**,** 1335-1344.

CROW, J. F. (1997) The high spontaneous mutation rate: is it a health risk? *Proc Natl Acad Sci USA,* 94**,** 8380-8386.

D'ONOFRIO, G., MOUCHIROUD, D., AÏSSANI, B., GAUTIER, C. & BERNARDI, G. (1991) Correlations Between the Compositional Properties of Human Genes, Codon Usage, and Amino Acid Composition of Proteins. *Journal of Molecular Evolution,* 32**,** 504-510.

DIMITROVA, D. S. & GILBERT, D. M. (1999) The spatial position and replication timing of chromosomal domains are both established in early G1 phase. *Molecular Cell,* 4**,** 983-993.

DRESZER, T. R., WALL, G. D., HAUSSLER, D. & POLLARD, K. S. (2007) Biased clustered substitutions in the human genome: the footprints of male-driven biased gene conversion. *Genome Research,* 17**,** 1420-1430.

DUMONT, B. L., WHITE, M. A., STEFFY, B., WILTSHIRE, T. & PAYSEUR, B. A. (2011) Extensive recombination rate variation in the house mouse species complex inferred from genetic linkage maps. *Genome Research,* 21**,** 114-125.

DURET, L. & ARNDT, P. F. (2008) The impact of recombination on nucleotide substitutions in the human genome. *PLoS Genet,* 4**,** e1000071.

DURET, L. & GALTIER, N. (2009) Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu Rev Genomics Hum Genet,* 10**,** 285-311.

DURET, L., SEMON, M., PIGANEAU, G., MOUCHIROUD, D. & GALTIER, N. (2002) Vanishing GC-rich isochores in mammalian genomes. *Genetics,* 162**,** 1837-1847.

ELLEGREN, H. (2007) Characteristics, causes and evolutionary consequences of male-biased mutation. *Proc Biol Sci,* 274**,** 1-10.

ELLEGREN, H. & FRIDOLFSSON, A.-K. (2003) Sex-specific mutation rates in salmonid fish. *Journal of Molecular Evolution,* 56**,** 458-463.

ELLISON, J., LI, X., FRANCKE, U. & SHAPIRO, L. (1996) Rapid evolution of human pseudoautosomal genes and their mouse homologs. *Mammalian Genome,* 7**,** 25-30.

EYRE-WALKER, A. (1993) Recombination and mammalian genome evolution. *Proc Biol Sci,* 252**,** 237-243.

EYRE-WALKER, A. & HURST, L. D. (2001) The evolution of isochores. *Nat Rev Genet,* 2**,** 549-555.

FARKASH-AMAR, S., LIPSON, D., POLTEN, A., GOREN, A., HELMSTETTER, C., YAKHINI, Z. & SIMON, I. (2008) Global organization of replication time zones of the mouse genome. *Genome Research,* 18**,** 1562-1570.

FARKASH-AMAR, S. & SIMON, I. (2009) Genome-wide analysis of the replication program in mammals. *Chromosome Res,* 18**,** 115-125.

FILATOV, D. A. (2004) A gradient of silent substitution rate in the human pseudoautosomal region. *Mol Biol Evol,* 21**,** 410-417.

FILATOV, D. A. & GERRARD, D. T. (2003) High mutation rates in human and ape pseudoautosomal genes. *Gene,* 317**,** 67-77.

FILIPSKI, J. (1987) Correlation between molecular clock ticking, codon usage fidelity of DNA repair, chromosome banding and chromatin compactness in germline cells. *FEBS Lett,* 217**,** 184-186.

GAFFNEY, D. J. & KEIGHTLEY, P. D. (2005) The scale of mutational variation in the murid genome. *Genome Research,* 15**,** 1086-1094.

GALTIER, N. (2004) Recombination, GC-content and the human pseudoautosomal boundary paradox. *Trends Genet,* 20**,** 347-349.

GALTIER, N. & DURET, L. (2007) Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. *Trends Genet,* 23**,** 273-277.

GALTIER, N., DURET, L., GLÉMIN, S. & RANWEZ, V. (2009) GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates. *Trends Genet,* 25**,** 1-5.

GILBERT, D. M. (2002) Replication timing and transcriptional control: beyond cause and effect. *Curr Opin Cell Biol,* 14**,** 377-383.

GOJOBORI, T., LI, W. & GRAUR, D. (1982) Patterns of Nucleotide Substitution in Pseudogenes and Functional Genes. *Journal of Molecular Evolution,* 18**,** 360-369.

GORIELY, A., MCVEAN, G. A. T., RÖJMYR, M., INGEMARSSON, B. & WILKIE, A. O. M. (2003) Evidence for selective advantage of pathogenic FGFR2 mutations in the male germ line. *Science,* 301**,** 643-646.

GORIELY, A., MCVEAN, G. A. T., VAN PELT, A. M. M., O'ROURKE, A. W., WALL, S. A., DE ROOIJ, D. G. & WILKIE, A. O. M. (2005) Gain-of-function amino acid substitutions drive positive selection of FGFR2 mutations in human spermatogonia. *Proc Natl Acad Sci USA,* 102**,** 6051-6056.

GREEN, P., EWING, B., MILLER, W., THOMAS, P. J., PROGRAM, N. C. S. & GREEN, E. D. (2003) Transcription-associated mutational asymmetry in mammalian evolution. *Nat Genet,* 33**,** 514-517.

GU, X. & LI, W. (1994) A Model for the Correlation of Mutation-Rate with GC Content and The Origin of GC-Rich Isochores. *Journal of Molecular Evolution,* 38**,** 468-475.

HALDANE, J. B. S. (1947) The mutation rate of the gene for hemophilia and its segregation ratios in males and females. *Ann. Eugen.,* 13**,** 262-271.

HARDISON, R. C., ROSKIN, K. M., YANG, S., DIEKHANS, M., KENT, W. J., WEBER, R., ELNITSKI, L., LI, J., O'CONNOR, M., KOLBE, D., SCHWARTZ, S., FUREY, T. S., WHELAN, S., GOLDMAN, N., SMIT, A., MILLER, W., CHIAROMONTE, F. & HAUSSLER, D. (2003) Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. *Genome Research,* 13**,** 13-26.

HELLMANN, I., PRÜFER, K., JI, H., ZODY, M. C., PÄÄBO, S. & PTAK, S. E. (2005) Why do human diversity levels vary at a megabase scale? *Genome Research,* 15**,** 1222-1231.

HINCH, A. G., TANDON, A., PATTERSON, N., SONG, Y., ROHLAND, N., PALMER, C. D., CHEN, G. K., WANG, K., BUXBAUM, S. G., AKYLBEKOVA, E. L., *et al.* (2011) The landscape of recombination in African Americans. *Nature,* 476**,** 170-175.

HIRATANI, I., LESKOVAR, A. & GILBERT, D. M. (2004) Differentiation-induced replication-timing changes are restricted to AT-rich/long interspersed nuclear element (LINE)-rich isochores. *Proc Natl Acad Sci USA,* 101**,** 16861-16866.

HIRATANI, I., RYBA, T., ITOH, M., YOKOCHI, T., SCHWAIGER, M., CHANG, C.-W., LYOU, Y., TOWNES, T. M., SCHÜBELER, D. & GILBERT, D. M. (2008) Global reorganization of replication domains during embryonic stem cell differentiation. *PLoS Biol,* 6**,** e245.

HODGKINSON, A. & EYRE-WALKER, A. (2011) Variation in the mutation rate across mammalian genomes. *Nature Reviews Genetics,* 12, 756-766.

HOLMQUIST, G. P. (1987) Role of replication time in the control of tissue-specific gene expression. *Am J Hum Genet,* 40**,** 151-173.

HURLES, M. E. (2001) Gene Conversion. *eLS.* John Wiley & Sons, Ltd.

HURST, L. D. & ELLEGREN, H. (1998) Sex biases in the mutation rate. *Trends Genet,* 14**,** 446-452.

HURST, L. D. & WILLIAMS, E. J. (2000) Covariation of GC content and the silent site substitution rate in rodents: implications for methodology and for the evolution of isochores. *Gene,* 261**,** 107-114.

JACKSON, D. A. & POMBO, A. (1998) Replicon clusters are stable units of chromosome structure: evidence that nuclear organization contributes to the efficient activation and propagation of S phase in human cells. *The Journal of Cell Biology,* 140**,** 1285-1295.

JEFFREYS, A. J. & NEUMANN, R. (2009) The rise and fall of a human recombination hot spot. *Nat Genet,* 41**,** 625-629.

JENSEN-SEAMAN, M. I., FUREY, T. S., PAYSEUR, B. A., LU, Y., ROSKIN, K. M., CHEN, C.-F., THOMAS, M. A., HAUSSLER, D. & JACOB, H. J. (2004) Comparative recombination rates in the rat, mouse, and human genomes. *Genome Research,* 14**,** 528-538.

KIMURA, M. (1968) Evolutionary rate at the molecular level. *Nature,* 217**,** 624-626.

KOÇ, A., WHEELER, L. J., MATHEWS, C. K. & MERRILL, G. F. (2004) Hydroxyurea arrests DNA replication by a mechanism that preserves basal dNTP pools. *J Biol Chem,* 279**,** 223-230.

KONG, A., GUDBJARTSSON, D. F., SAINZ, J., JONSDOTTIR, G. M., GUDJONSSON, S. A., RICHARDSSON, B., SIGURDARDOTTIR, S., BARNARD, J., HALLBECK, B., MASSON, G., SHLIEN, A., PALSSON, S. T., FRIGGE, M. L., THORGEIRSSON, T. E., GULCHER, J. R. & STEFANSSON, K. (2002) A high-resolution recombination map of the human genome. *Nat Genet,* 31**,** 241-247.

KONG, A., THORLEIFSSON, G., GUDBJARTSSON, D. F., MASSON, G., SIGURDSSON, A., JONASDOTTIR, A., WALTERS, G. B., JONASDOTTIR, A., GYLFASON, A., KRISTINSSON, K. T., GUDJONSSON, S. A., FRIGGE, M. L., HELGASON, A., THORSTEINSDOTTIR, U. & STEFANSSON, K. (2010) Fine-scale

recombination rate differences between sexes, populations and individuals. *Nature,* 467**,** 1099-1103.
KULATHINAL, R. J., BENNETT, S. M., FITZPATRICK, C. L. & NOOR, M. A. F. (2008) Fine-scale mapping of recombination rate in *Drosophila* refines its correlation to diversity and divergence. *Proc Natl Acad Sci USA,* 105**,** 10051-10056.
KUMAR, D., ABDULOVIC, A. L., VIBERG, J., NILSSON, A. K., KUNKEL, T. A. & CHABES, A. (2011) Mechanisms of mutagenesis in vivo due to imbalanced dNTP pools. *Nucleic Acids Res,* 39**,** 1360-1371.
KUMAR, D., VIBERG, J., NILSSON, A. K. & CHABES, A. (2010) Highly mutagenic and severely imbalanced dNTP pools can escape detection by the S-phase checkpoint. *Nucleic Acids Res,* 38**,** 3975-3983.
LANGLEY, C. H., LAZZARO, B. P., PHILLIPS, W., HEIKKINEN, E. & BRAVERMAN, J. M. (2000) Linkage disequilibria and the site frequency spectra in the su(s) and su(w(a)) regions of the *Drosophila melanogaster* X chromosome. *Genetics,* 156**,** 1837-1852.
LEEDS, J. M., SLABAUGH, M. B. & MATHEWS, C. K. (1985) DNA precursor pools and ribonucleotide reductase activity: distribution between the nucleus and cytoplasm of mammalian cells. *Mol Cell Biol,* 5**,** 3443-3450.
LENORMAND, T. (2003) The evolution of sex dimorphism in recombination. *Genetics,* 163**,** 811-822.
LENORMAND, T. & DUTHEIL, J. (2005) Recombination difference between sexes: a role for haploid selection. *PLoS Biol,* 3**,** e63.
LERCHER, M. J., CHAMARY, J.-V. & HURST, L. D. (2004) Genomic regionality in rates of evolution is not explained by clustering of genes of comparable expression profile. *Genome Research,* 14**,** 1002-1013.
LERCHER, M. J. & HURST, L. D. (2002) Human SNP variability and mutation rate are higher in regions of high recombination. *Trends Genet,* 18**,** 337-340.
LERCHER, M. J. & HURST, L. D. (2003) Imprinted chromosomal regions of the human genome have unusually high recombination rates. *Genetics,* 165**,** 1629-1632.
LERCHER, M. J., WILLIAMS, E. J. & HURST, L. D. (2001) Local similarity in evolutionary rates extends over whole chromosomes in human-rodent and mouse-rat comparisons: implications for understanding the mechanistic basis of the male mutation bias. *Mol Biol Evol,* 18**,** 2032-2039.
LIEN, S., SZYDA, J., SCHECHINGER, B., RAPPOLD, G. & ARNHEIM, N. (2000) Evidence for heterogeneity in recombination in the human pseudoautosomal region: high resolution analysis by sperm typing and radiation-hybrid mapping. *Am J Hum Genet,* 66**,** 557-566.
MACALPINE, D. M. & BELL, S. P. (2005) A genomic view of eukaryotic DNA replication. *Chromosome Res,* 13**,** 309-326.
MAGNI, G. (1963) The origin of spontaneous mutations during meiosis. *Proc Natl Acad Sci USA,* 50**,** 975-980.
MAJEWSKI, J. (2003) Dependence of mutational asymmetry on gene-expression levels in the human genome. *Am J Hum Genet,* 73**,** 688-692.
MAKOVA, K. D. & LI, W.-H. (2002) Strong male-driven evolution of DNA sequences in humans and apes. *Nature,* 416**,** 624-626.
MAKOVA, K. D., YANG, S. & CHIAROMONTE, F. (2004) Insertions and deletions are male biased too: a whole-genome analysis in rodents. *Genome Research,* 14**,** 567-573.

MALCOM, C. M., WYCKOFF, G. J. & LAHN, B. T. (2003) Genic mutation rates in mammals: local similarity, chromosomal heterogeneity, and X-versus-autosome disparity. *Mol Biol Evol,* 20**,** 1633-1641.

MALÍNSKY, J., KOBERNA, K., STANĚK, D., MASATA, M., VOTRUBA, I. & RASKA, I. (2001) The supply of exogenous deoxyribonucleotides accelerates the speed of the replication fork in early S-phase. *J Cell Sci,* 114**,** 747-750.

MARTOMO, S. A. & MATHEWS, C. K. (2002) Effects of biological DNA precursor pool asymmetry upon accuracy of DNA replication in vitro. *Mutat Res,* 499**,** 197-211.

MATASSI, G., SHARP, P. M. & GAUTIER, C. (1999) Chromosomal location effects on gene sequence evolution in mammals. *Curr Biol,* 9**,** 786-791.

MATHEWS, C. K. (2006) DNA precursor metabolism and genomic stability. *FASEB J,* 20**,** 1300-1314.

MCVEAN, G. & MYERS, S. (2010) PRDM9 marks the spot. *Nat Genet,* 42**,** 821-822.

MCVEAN, G. T. & HURST, L. D. (1997) Evidence for a selectively favourable reduction in the mutation rate of the X chromosome. *Nature,* 386**,** 388-392.

MEUNIER, J. & DURET, L. (2004) Recombination drives the evolution of GC-content in the human genome. *Mol Biol Evol,* 21**,** 984-990.

MIYATA, T., HAYASHIDA, H., KUMA, K., MITSUYASU, K. & YASUNAGA, T. (1987a) Male-driven Molecular Evolution: A Model and Nucleotide Sequence Analysis. *Cold Spring Harbor Symposia on Quantitative Biology,* 52**,** 863-867.

MIYATA, T., HAYASHIDA, H., KUMA, K. & YASUNAGA, T. (1987b) Male-driven molecular evolution demonstrated by different rates of silent substitutions between autosome-and sex chromosome-linked genes. *Proceedings of the Japan Academy. Ser. B: Physical and Biological Sciences,* 63**,** 327-331.

MUGAL, C. F., VON GRÜNBERG, H.-H. & PEIFER, M. (2009) Transcription-induced mutational strand bias and its effect on substitution rates in human genes. *Mol Biol Evol,* 26**,** 131-142.

MYERS, S., BOWDEN, R., TUMIAN, A., BONTROP, R. E., FREEMAN, C., MACFIE, T. S., MCVEAN, G. & DONNELLY, P. (2010) Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. *Science,* 327**,** 876-879.

MYERS, S., FREEMAN, C., AUTON, A., DONNELLY, P. & MCVEAN, G. (2008) A common sequence motif associated with recombination hot spots and genome instability in humans. *Nat Genet,* 40**,** 1124-1129.

MYERS, S., SPENCER, C. C. A., AUTON, A., BOTTOLO, L., FREEMAN, C., DONNELLY, P. & MCVEAN, G. (2006) The distribution and causes of meiotic recombination in the human genome. *Biochem Soc Trans,* 34**,** 526-530.

NACHMAN, M. W. (2001) Single nucleotide polymorphisms and recombination rate in humans. *Trends Genet,* 17**,** 481-485.

NAKAGOME, S., PECON-SLATTERY, J. & MASUDA, R. (2008) Unequal rates of Y chromosome gene divergence during speciation of the family Ursidae. *Mol Biol Evol,* 25**,** 1344-1356.

OLIVER, P. L., GOODSTADT, L., BAYES, J. J., BIRTLE, Z., ROACH, K. C., PHADNIS, N., BEATSON, S. A., LUNTER, G., MALIK, H. S. &

PONTING, C. P. (2009) Accelerated evolution of the Prdm9 speciation gene across diverse metazoan taxa. *PLoS Genet,* 5**,** e1000753.
PAIGEN, K., SZATKIEWICZ, J. P., SAWYER, K., LEAHY, N., PARVANOV, E. D., NG, S. H. S., GRABER, J. H., BROMAN, K. W. & PETKOV, P. M. (2008) The recombinational anatomy of a mouse chromosome. *PLoS Genet,* 4**,** e1000119.
PARVANOV, E. D., PETKOV, P. M. & PAIGEN, K. (2010) Prdm9 controls activation of mammalian recombination hotspots. *Science,* 327**,** 835.
PERRY, J. & ASHWORTH, A. (1999) Evolutionary rate of a gene affected by chromosomal position. *Curr Biol,* 9**,** 987-989.
PETKOV, P. M., BROMAN, K. W., SZATKIEWICZ, J. P. & PAIGEN, K. (2007) Crossover interference underlies sex differences in recombination rates. *Trends in Genetics,* 23**,** 539-542.
POLAK, P. & ARNDT, P. F. (2008) Transcription induces strand-specific mutations at the 5' end of human genes. *Genome Research,* 18**,** 1216-1223.
PRENDERGAST, J. G. D., CAMPBELL, H., GILBERT, N., DUNLOP, M. G., BICKMORE, W. A. & SEMPLE, C. A. M. (2007) Chromatin structure and evolution in the human genome. *BMC Evol Biol,* 7**,** 72.
PTAK, S., HINDS, D., KOEHLER, K., NICKEL, B., PATIL, N., BALLINGER, D., PRZEWORSKI, M., FRAZER, K. & PÄÄBO, S. (2005) Fine-scale recombination patterns differ between chimpanzees and humans. *Nat Genet,* 37**,** 429-434.
QIN, J., CALABRESE, P., TIEMANN-BOEGE, I., SHINDE, D. N., YOON, S.-R., GELFAND, D., BAUER, K. & ARNHEIM, N. (2007) The molecular anatomy of spontaneous germline mutations in human testes. *PLoS Biol,* 5**,** e224.
RAGHURAMAN, M. K., WINZELER, E. A., COLLINGWOOD, D., HUNT, S., WODICKA, L., CONWAY, A., LOCKHART, D. J., DAVIS, R. W., BREWER, B. J. & FANGMAN, W. L. (2001) Replication dynamics of the yeast genome. *Science,* 294**,** 115-121.
RATTRAY, A. J. & STRATHERN, J. N. (2003) Error-prone DNA polymerases: when making a mistake is the only way to get ahead. *Annu Rev Genet,* 37**,** 31-66.
ROMIGUIER, J., RANWEZ, V., DOUZERY, E. J. P. & GALTIER, N. (2010) Contrasting GC-content dynamics across 33 mammalian genomes: Relationship with life-history traits and chromosome sizes. *Genome Research,* 20**,** 1001-1009.
RYBA, T., HIRATANI, I., SASAKI, T., BATTAGLIA, D., KULIK, M., ZHANG, J., DALTON, S. & GILBERT, D. M. (2011) Replication timing: a fingerprint for cell identity and pluripotency. *PLoS Comput Biol,* 7**,** e1002225.
SANDSTEDT, S. A. & TUCKER, P. K. (2005) Male-driven evolution in closely related species of the mouse genus *Mus*. *Journal of Molecular Evolution,* 61**,** 138-144.
SAYRES, M. A. W., VENDITTI, C., PAGEL, M. & MAKOVA, K. D. (2011) Do Variations in Substitution Rates and Male Mutation Bias Correlate With Life-History Traits? A Study of 32 Mammalian Genomes. *Evolution,* 65**,** 2800-2815.
SCHMEGNER, C., HAMEISTER, H., VOGEL, W. & ASSUM, G. (2007) Isochores and replication time zones: a perfect match. *Cytogenetic and Genome Research,* 116**,** 167-172.

SCHÜBELER, D., SCALZO, D., KOOPERBERG, C., VAN STEENSEL, B., DELROW, J. & GROUDINE, M. (2002) Genome-wide DNA replication profile for *Drosophila melanogaster*: a link between transcription and replication timing. *Nat Genet,* 32**,** 438-442.

SCHWAIGER, M., STADLER, M. B., BELL, O., KOHLER, H., OAKELEY, E. J. & SCHÜBELER, D. (2009) Chromatin state marks cell-type- and gender-specific replication of the *Drosophila* genome. *Genes Dev,* 23**,** 589-601.

SELIG, S., OKUMURA, K., WARD, D. C. & CEDAR, H. (1992) Delineation of DNA replication time zones by fluorescence in situ hybridization. *The EMBO Journal,* 11**,** 1217-1225.

SHIFMAN, S., BELL, J. T., COPLEY, R. R., TAYLOR, M. S., WILLIAMS, R. W., MOTT, R. & FLINT, J. (2006) A high-resolution single nucleotide polymorphism genetic map of the mouse genome. *PLoS Biol,* 4**,** e395.

SHIMMIN, L. C., CHANG, B. H. & LI, W. H. (1993) Male-driven evolution of DNA sequences. *Nature,* 362**,** 745-747.

SMITH, N. G. & HURST, L. D. (1999) The causes of synonymous rate variation in the rodent genome. Can substitution rates be used to estimate the sex bias in mutation rate? *Genetics,* 152**,** 661-673.

SORIANO, P., KEITGES, E. A., SCHORDERET, D. F., HARBERS, K., GARTLER, S. M. & JAENISCH, R. (1987) High rate of recombination and double crossovers in the mouse pseudoautosomal region during male meiosis. *Proc Natl Acad Sci USA,* 84**,** 7218-7220.

STRATHERN, J. N., SHAFER, B. K. & MCGILL, C. B. (1995) DNA synthesis errors associated with double-strand-break repair. *Genetics,* 140**,** 965-972.

SUNDSTRÖM, H., WEBSTER, M. T. & ELLEGREN, H. (2003) Is the rate of insertion and deletion mutation male biased?: Molecular evolutionary analysis of avian and primate sex chromosome sequences. *Genetics,* 164**,** 259-268.

TEASE, C. & HULTEN, M. (2004) Inter-sex variation in synaptonemal complex lengths largely determine the different recombination rates in male and female germ cells. *Cytogenetic and Genome Research,* 107**,** 208-215.

TIAN, D., WANG, Q., ZHANG, P., ARAKI, H., YANG, S., KREITMAN, M., NAGYLAKI, T., HUDSON, R., BERGELSON, J. & CHEN, J.-Q. (2008) Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes. *Nature,* 455**,** 105-108.

TICHER, A. & GRAUR, D. (1989) Nucleic-Acid Composition, Codon Usage, and the Rate of Synonymous Substitution in Protein-Coding Genes. *Journal of Molecular Evolution,* 28**,** 286-298.

TOUCHON, M., ARNEODO, A., D'AUBENTON-CARAFA, Y. & THERMES, C. (2004) Transcription-coupled and splicing-coupled strand asymmetries in eukaryotic genomes. *Nucleic Acids Res,* 32**,** 4969-4978.

TOUCHON, M., NICOLAY, S., ARNEODO, A., D'AUBENTON-CARAFA, Y. & THERMES, C. (2003) Transcription-coupled TA and GC strand asymmetries in the human genome. *FEBS Lett,* 555**,** 579-582.

TYEKUCHEVA, S., MAKOVA, K. D., KARRO, J. E., HARDISON, R. C., MILLER, W. & CHIAROMONTE, F. (2008) Human-macaque comparisons illuminate variation in neutral substitution rates. *Genome Biol,* 9**,** R76.

WILLIAMS, E. J. & HURST, L. D. (2000) The proteins of linked genes evolve at similar rates. *Nature,* 407**,** 900-903.

WOLFE, K. H. & SHARP, P. M. (1993) Mammalian Gene Evolution - Nucleotide-Sequence Divergence Between Mouse and Rat. *Journal of Molecular Evolution,* 37**,** 441-456.

WOLFE, K. H., SHARP, P. M. & LI, W. H. (1989) Mutation rates differ among regions of the mammalian genome. *Nature,* 337**,** 283-285.

WOODFINE, K., FIEGLER, H., BEARE, D. M., COLLINS, J. E., MCCANN, O. T., YOUNG, B. D., DEBERNARDI, S., MOTT, R., DUNHAM, I. & CARTER, N. P. (2004) Replication timing of the human genome. *Hum Mol Genet,* 13**,** 191-202.

WU, Z. K., GETUN, I. V. & BOIS, P. R. J. (2010) Anatomy of mouse recombination hot spots. *Nucleic Acids Res,* 38**,** 2346-2354.

XU, Y. Z., HUANG, P. & PLUNKETT, W. (1995) Functional compartmentation of dCTP pools. Preferential utilization of salvaged deoxycytidine for DNA repair in human lymphoblasts. *J Biol Chem,* 270**,** 631-637.

YAFFE, E., FARKASH-AMAR, S., POLTEN, A., YAKHINI, Z., TANAY, A. & SIMON, I. (2010) Comparative analysis of DNA replication timing reveals conserved large-scale chromosomal architecture. *PLoS Genet,* 6**,** e1001011.

YI, S., SUMMERS, T. J., PEARSON, N. M. & LI, W.-H. (2004) Recombination has little effect on the rate of sequence divergence in pseudoautosomal boundary 1 among humans and great apes. *Genome Research,* 14**,** 37-43.

Chapter 2. **Evidence that Replication-Associated Mutation Alone Does Not Explain Between-Chromosome Differences in Substitution Rates**

Catherine J. Pink, Siva K. Swaminathan, Ian Dunham, Jane Rogers, Andrew Ward, and Laurence D. Hurst

Based on a paper and supplementary information published at:

Genome Biology and Evolution (2009). 1(1): 13-22

2.1 Introduction

Following Haldane's (1947) proposal that most mutations in humans are male derived, it has been conventional wisdom that this male excess is owing to a difference in numbers of germ-line replications (Crow, 1997a, Crow, 1997b, Hurst and Ellegren, 1998, Ellegren, 2007, Li *et al*., 2002). In males, spermatogenesis is an ongoing process throughout a male's life, whereas in females, the number of divisions prior to oocyte production is fixed. Under the presumption of a male bias owing to replication number differences, Miyata *et al*. (1987) proposed a simple means to assay the extent of male-bias. They argued that the rate of evolution of putatively neutral sites on X, Y and autosomes should reflect the amount of time spent in the male germ-line by the three chromosomal classes: The Y chromosome should evolve the fastest being exclusively in males, followed by the autosomes that are, on average, in males half of the time, followed by the X chromosome which spends only one third of its time in males.

More formally, Miyata *et al*. (1987) proposed that if the mutation rate in females is μ and the ratio of male-to-female germ-line replication events (prior to generation of a successful gamete) is α, then if germ-line replication is the source of all differences

in substitution rate of sequence not under selection, the evolutionary rate (K) of sequences of class N (Y, X or autosomal) would be:

$$K_Y = \alpha\mu, \tag{1}$$

$$K_{Autosome} = \frac{\alpha\mu + \mu}{2}, \tag{2}$$

$$K_X = \frac{\alpha\mu + 2\mu}{3} \tag{3}$$

By considering the ratios of any two classes at a time ($K_X/K_{Autosome}$, $K_Y/K_{Autosome}$ or K_Y/K_X), it is possible to estimate α.

Typically, by employing just one of the three possible comparisons, various authors have attempted to assess the extent of male bias in various taxa (eg. Shimmin *et al*., 1993, Chang *et al*., 1994, Makova and Li, 2002, Sandstedt and Tucker, 2005, Goetting-Minesky and Makova, 2006, Bachtrog, 2008). It is commonly argued (Makova and Li, 2002) that results are broadly consistent with expectations, in that species with relatively long-lived males (hence, a greater discrepancy between the number of male and female replications) have higher values of α. For humans, the estimate is typically around six (Shimmin *et al*., 1993, Taylor *et al*., 2006, Chang *et al*., 1996), for rodents around two (Chang *et al*., 1994, Sandstedt and Tucker, 2005) and for flies around one (Bauer and Aquadro, 1997).

The case is, however, by no means decided. First, direct observations of male-bias derived from medical evidence, rather than molecular evolutionary inferred estimates, do not agree with one another (Hurst and Ellegren, 1998, Hurst, 2006). In part, this reflects the fact that very high estimates of α appear to be confounded by male germ-line selection favouring certain mutations (Goriely *et al*., 2003, Goriely *et al*., 2005, Qin *et al*., 2007, Choi *et al*., 2008). Although these very strong male-biases were initially taken as strong support for the replication hypothesis (Li *et al*., 2002), they no longer arbitrate on the issue. Why some studies (e.g. Yin *et al*., 1996) might show a female bias, while the direction of the bias is family specific in others (Conrad *et al*., 2011) is unresolved.

The molecular evolutionary comparisons also have a number of unresolved issues. Z-W chromosomal comparisons in birds, for example, tend to give estimates

(Bartosch-Härlid *et al*., 2003) that are rather high given the short life spans of the species ($\alpha \sim 5$) (Hurst and Ellegren, 1998). In *Drosophila*, one study claims there is a bias of the same magnitude claimed for rodents (Bachtrog, 2008). In mammals there is also now strong evidence for within-autosome (Matassi *et al*., 1999, Lercher *et al*., 2001, Malcom *et al*., 2003) and between-autosome (Lercher *et al*., 2001, Malcom *et al*., 2003) variation in rates that cannot be accounted for by differences in the number of replications, this being the same across all autosomes.

Although it has been claimed that differences in the rate of evolution of different chromosomal classes can be explained by the male mutation bias alone (Axelsson *et al*., 2004), others have argued that mutations arising in non-replicating DNA also contribute substantially to rates of evolution (Huttley *et al*., 2000). Further, substitution rates are known to be effected by transcription (Green *et al*., 2003, Majewski, 2003, Lercher *et al*., 2004), location within an inversion (Navarro and Barton, 2003), GC content (Smith and Hurst, 1999b, Hurst and Williams, 2000) and recombination (Perry and Ashworth, 1999, Dreszer *et al*., 2007, Bussell *et al*., 2006, Rattray *et al*., 2001, Lercher and Hurst, 2002, Hellmann *et al*., 2003). However, the quantitative effect of these processes, if any, on Miyata *et al*.'s (1987) model has not yet been explored.

This chapter aims to provide an examination of the model proposed by Miyata *et al*. (1987). There is a simple test of whether replication alone is the source of the differences in evolutionary rate between X, Y and autosome. If the model is correct, then Equations 1 to 3 must hold. If so, all of the possible pair-wise comparisons (X-Autosome, Y-Autosome and Y-X) should provide the same estimate for α. If they do not, then the "replication-number alone" method fails and application of Miyata *et al*.'s (1987) commonly employed method must be questioned.

In rodents, some prior evidence suggests that the value of α is dependent on which chromosomal classes were employed (Smith and Hurst, 1999a). However, sample sizes were too limited to make definitive statements and substitution rates at exonic silent sites were used. As it is now known (Chamary *et al*., 2006) that selection can act on synonymous mutations in mammals (although estimates of K_S are very similar

to K_i – the intronic rate), it was worthwhile repeating this analysis using a larger sample of well-aligned intronic sequence as well as employing synonymous rates.

2.2 Methods

2.2.1 Sequences

Mouse (*Mus musculus*) and rat (*Rattus norvegicus*) autosomal and X-linked sequences were downloaded from the University of California Santa Cruz (UCSC) Genome Bioinformatics database (Karolchik *et al*., 2004, www.genome.ucsc.edu). Mouse exonic and intronic sequences were obtained from the February 2006 and July 2007 builds respectively, whereas rat exonic and intronic sequences were both obtained from the November 2004 build. Exonic sequences were concatenated by gene and filtered so that only those containing complete codons, correct start and termination codons, no premature stops, or ambiguous bases were retained. All sequences pertaining to genes that failed this test were removed from the data set.

The completed genome sequence of the Brown Norway rat did not include the Y chromosome (Gibbs *et al*., 2004). In order to provide the Y-linked sequence necessary for this analysis, Dr Swaminathan sequenced two rat Y-linked bacterial artificial chromosomes (BACs) plus some additional Y-linked cDNAs, the latter intended to expand the inventory for analysis of synonymous substitution rates. The methodologies used for this sequencing are detailed in Pink *et al*. (2009, Appendix 2). Although copious amounts of sequence were produced, outside of the coding regions it proved impossible for Dr Swaminathan to unambiguously define orthology. The analysis of BAC derived sequence was therefore confined to well-aligned genic sequences.

This rat Y-linked sequencing gave rise to the full intronic sequences of *Ube1y* and *Eif2s3y* and a partial intronic sequence of *Jarid1d*. The last intron of *Zfy* was obtained from accession file X58934. A blastn search of rat Y-linked sequences against the mouse genome identified orthologous Y-linked mouse genes, for which intronic sequences were downloaded from the UCSC Genome Bioinformatics database (Karolchik *et al*., 2004). In addition, full coding sequences of rat *Ube1y* and

Eif2s3y and partial coding sequences of *Ddx3y* (alias *Dby*), *Uty* and *Jarid1d* (alias *Smcy*) were obtained. Extraction and re-formatting of these Y-linked sequences ready for analysis was performed by Dr Swaminathan and Dr Batada.

The correct reading frame of the partial exonic sequences was established as that free from internal stop codons. The full coding sequence of rat *Sry* and partial coding sequence of rat *Zfy* were obtained from accession files NM_012772 and X75172 respectively. Rat Y-linked sequences were subjected to a blastn search against the mouse genome to identify orthologous mouse Y-linked coding sequence, for which sequences were obtained from GenBank. Full details of the sequences used are given in Table 2.1.

	Intronic			**Exonic**		
	Sequence	**Rat ID**	**Mouse ID**	**Sequence**	**Rat ID**	**Mouse ID**
Ube1y	Full	NA 17060 bp	NM_011667 19411 bp	Full	FJ775730 NM_001167666 3177 bp	NM_011667 3980 bp
Eif2s3y	Full	NA 5889 bp	NM_012011 6225 bp	Full	FJ775731 FJ775732 1419 bp	NM_012011 1801 bp
Ddx3y	-	-	-	Partial	FJ775727 NM_001167665 1207 bp	NM_012008 3767 bp
Uty	-	-	-	Partial	FJ775728 3283 bp	NM_009484 3823 bp
Jarid1d	Partial	NA 13065 bp	NM_011419 13265 bp	Partial	FJ775729 2910 bp	NM_011419 5478 bp
Sry	-	-	-	Full	NM_012772 510 bp	NM_011564 1188 bp
Zfy	Last intron	X58934 994 bp	NM_009570 913 bp	Partial	X75172 1173 bp	NM_009571 2794 bp

Table 2.1: Sources and lengths of all Y-linked sequences used for the analyses.

2.2.2 Ortholog identification

From an initial set of Mouse Genome Informatics (MGI)-defined mouse-rat orthologs (Eppig *et al*., 2007), downloaded from http://www.informatics.jax.org in February 2007, autosomal and X-linked orthologs were further strictly defined by reference to exon number and phase, mouse and rat having to be the same, and by genomic location, chromosomal class having to be known and of the same type. Intronic orthologs were further filtered to retain only those where the difference in coding sequence lengths was less than 5% of the mean coding sequence length.

Orthologous Y-linked genes were identified from blastn search of rat sequence against the mouse genome as previously described.

2.2.3 Alignments

40,168 orthologous introns were aligned individually using LAGAN (Brudno *et al*., 2003), with exons identified by reference to mouse and/or rat cDNA in the case of new Y-linked sequence or by RefSeq annotation in the case of X-linked and autosomal genes.

By reference to a set of hand-aligned mouse-rat introns, Chamary and Hurst (2004) determined that there should be no more than 0.84 indels per base pair of alignment and that the alignment length should be no greater than 1.16 times the length of the longest sequence. In all, 1,915 introns were eliminated due to failing to meet these criteria. Rates of evolution of autosomal introns derived from the LAGAN alignment ($K_i = 0.1666$) were in agreement with those previously obtained from introns aligned both manually ($K_i = 0.1533$) and using a maximum likelihood protocol ($K_i = 0.1791$) (Chamary and Hurst, 2004).

Coding sequences were concatenated by gene and their translations aligned using MUSCLE (Edgar, 2004) under default parameters, from which the nucleotide alignment was reconstructed. Exonic alignments of less than 300 sites, equivalent to 100 amino acids assuming no indels, were excluded from the analysis to control for bias introduced due to the influence of short sequences, these tending to be found in highly expressed and thus highly conserved genes (Castillo-Davis *et al*., 2002, Drummond *et al*., 2005).

2.2.4 Filter for introns with hidden exons or other constrained domains

Given the possibility of alternative splicing, it was possible, if not likely, that some of the above introns may have contained hidden exons, or indeed other residues under selection such as binding or regulatory domains. As introns containing regions under purifying selection and those experiencing low mutation rates would both exhibit low levels of divergence, it was not possible to use this alone as a criteria by

which non-neutrally evolving introns could be identified: the removal of introns with low levels of divergence would have introduced bias into the data set.

An alternative method to filter out these introns asked whether, within an intron, substitutions and conserved residues were clustered or randomly scattered through the intron. The premise was that if a hidden exon or a protein-binding domain was present, such regions should be relatively free of substitutions, so longer runs of conserved residues would be found, compared to what would be expected in the absence of such domains.

The filter consisted of a simulation in which varying percentages of diverged bases ranging from 10% to 90% (at 10% intervals) were randomly distributed along sequences varying in length from 100 to 100,000 bases. For each sequence length and percentage divergence modelled, the number of switches in state between conserved and diverged bases as one moved down the sequence was counted. For a given sequence length and percentage divergence, multiple permutations were run ranging from 10,000 permutations for shorter sequences to 100 permutations for longer sequences due to computational limitations. The number of switches in state for each permutation was ranked, from which the lowest one-sided 95 percentile was identified (Figure 2.1).

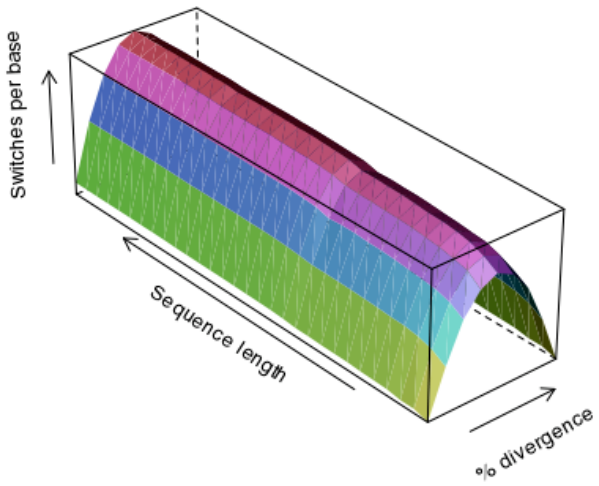


Figure 2.1: Based on the lowest one-sided 95 percentile, the lowest number of switches in state expected between conserved and diverged bases as one moves along a sequence, derived from a simulation in which a range of sequence lengths were randomly assigned varying percentage of diverged sites.

From this lower 95 percentile, a linear model was developed from which the lowest number of switches in state per base (z) expected for a given number of aligned nucleotides (l) and a given percentage of diverged bases (d) could be predicted by:

$$z = -0.005757 + 0.00000026(l) + 0.0192327(d) - 0.000192(d^2) + 0.0000000136((l - 20350)(d - 50)) - 0.00000000014((l - 20350)(d^2 - 3166.67)). \quad (4)$$

The plot of this model in Figure 2.2 shows how closely the linear model predicts the simulation shown in Figure 2.1. This method was not expected to be perfect (it was likely to miss small hidden exons), but it should have eliminated those introns most profoundly affected by hidden exons.

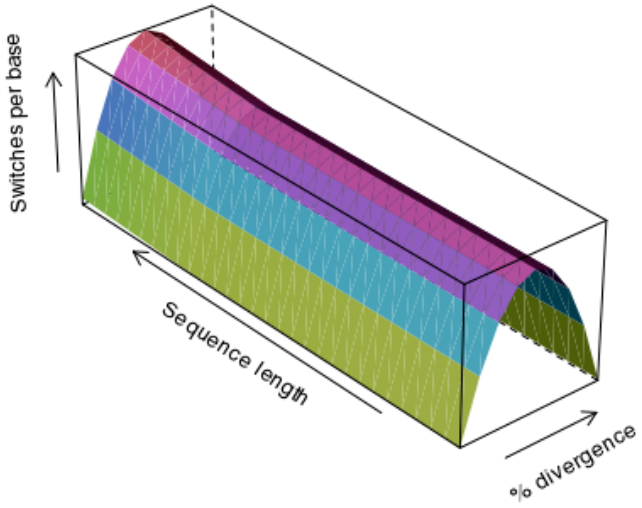


Figure 2.2: The lowest number of switches in state expected for a given number of aligned nucleotides and a given percentage of diverged bases, as predicted by a linear model.

After elimination of the 30 bp of sequence flanking exon-intron boundaries (known to be under selective constraint, Chamary and Hurst, 2004), sites were classified as conserved or diverged and then the number of switches in state as one moved down the intron was counted. By reference to the linear model, any intron showing a lower number of switches than predicted (z) was eliminated. In all, 21,041 (55%) introns showed such evidence of selective constraints and were excluded from the analysis. Autosomal rates of evolution remained largely unchanged between the purged and the unpurged data sets. Note that the filter removed one intron (that of *Zfy*)

previously employed in rodents to estimate the evolutionary rate of the Y chromosome.

2.2.5 Assignment of chromosomal location and concatenation of intronic sequences

The first intron of each gene was eliminated from the analysis, these known to be unusually slow evolving (Chamary and Hurst, 2004, Keightley and Gaffney, 2003). Indels were removed from the remaining intronic alignments. For estimation of chromosomal rates, all intronic sequence from a given chromosome, assigned by the location of the mouse ortholog, was concatenated. This intronic chromosomal data set consisted of 15,625 autosomal introns (16.7 Mb), 349 X-linked introns (450 Kb) and 20 Y-linked introns (6,624 bp).

For analysis of the effect of GC content, expression rate and a past history of inversions, introns from the same gene were concatenated (comprising 4051 autosomal, 107 X-linked and 3 Y-linked genes).

Aligned coding sequences for each gene were assigned to the three chromosomal classes, again based on the location of the mouse ortholog. This exonic data set was comprised of 4,474 autosomal genes (5.8 Mb), 145 X-linked genes (180 Kb) and 7 Y-linked genes (13662 bp).

2.2.6 Distance estimation

The rate of intronic divergence (K_i) was estimated and corrected for multiple hits according to the model of Tamura and Kumar (TK) (2002), this correcting for inhomogeneous evolution. This was used for the main analysis and all analyses at the genic level. Several other methods were additionally employed, including those of Jukes and Cantor (JC) (1969), Kimura (KM) (1980) and Tamura and Nei (TN) (1993).

K_S was estimated from exonic alignments using Li's (1993) protocol, correcting for multiple hits according to Kimura's two parameter model (Kimura, 1980). As methods for estimating the synonymous substitution rates are subject to

overestimation (McVean and Hurst, 1997), the synonymous rate at 4-fold degenerate sites (K_4) was also estimated, correcting for multiple hits according to Jukes and Cantor (1969), Kimura (1980) and Tamura and Nei (1993).

Exonic chromosomal *K* estimates were calculated from the average substitution rate of genes assigned to each chromosomal class. This was repeated using mean, mean weighted by alignment length and median measures of centrality.

Intronic X- and Y-linked substitution rates were determined directly from concatenated sequences assigned to each chromosome. Intronic autosomal substitution rates were calculated from the average substitution rate of each concatenated autosomal alignment. Analyses were repeated using the mean, mean weighted by alignment length and median measures of centrality. The main analyses utilised comparisons of intronic X- and Y-linked substitution rates to the autosomal mean. For analyses at the genic level, chromosomal means weighted by alignment length were used.

2.2.7 Estimation of α

The ratio of chromosomal *K* for each pairwise comparison (K_X to $K_{Autosome}$, K_Y to $K_{Autosome}$ and K_Y to K_X) were substituted into the equations of Miyata *et al*. (1987), namely

$$\alpha_{XAutosome} = \frac{\left(3\left(K_X / K_{Autosome}\right) - 4\right)}{\left(2 - 3\left(K_X / K_{Autosome}\right)\right)} \quad (5)$$

$$\alpha_{YAutosome} = \frac{\left(K_Y / K_{Autosome}\right)}{\left(2 - \left(K_Y / K_{Autosome}\right)\right)}, \quad (6)$$

$$\text{and } \alpha_{YX} = \frac{2\left(K_Y / K_X\right)}{\left(3 - \left(K_Y / K_X\right)\right)} \quad (7)$$

in order to calculate the male-to-female mutation rate ratio (α).

2.2.8 Error limits

95% Confidence intervals were determined via bootstrapping: Within each chromosomal class, per-gene synonymous substitution rates derived from coding sequences were randomly sampled, with replacement and preserving sample size,

from which an average substitution rate for the chromosomal type was determined, using each *K* estimator and measure of centrality previously described. Similarly, alignments of the same length as the concatenated intronic chromosomal sequences were created by random sampling of aligned intronic base pairs with replacement, from which chromosomal substitution rates were calculated using each *K* estimator previously described and average autosomal rates were determined using the three alternative measures of centrality. Substitution of these randomly sampled chromosomal rates of evolution for any given rate estimator into Equations 5 to 7 enabled estimation of α for each pair-wise comparison. Likewise for estimates of α, r and r_m derived from the two novel models. This process was repeated 10,000 times. These randomly generated estimates of each parameter were then ranked and the values lying at the 95 percentiles (i.e. at the 2.5 percentile and the 97.5 percentile) identified.

Significant differences between the rate of evolution of different chromosomal classes were determined from 10,000 permutations, whereby for each comparison, pairs of bootstrapped estimates were randomly sampled and the number of occasions on which either the estimates were equal or the chromosomal class with the higher rate was not that originally observed was counted, such that significance was calculated as P = (count + 1)/10,001. This method was also used to determine whether estimates of α were significantly different.

2.2.9 Recombination rates

Rat sex-averaged recombination rates over 5 Mb windows were obtained from Jensen-Seaman *et al*. (2004). These rates were derived from the physical position of markers placed on a previous build, RGSPC version 3.1 (rn3). To control for potential inaccuracies arising from incorrect annotation of these positions, the relative proximity of neighbouring genes in the previous build, RGSPC version 3.1, and the current build, RGSPC version 3.4 (rn4), were compared. A most conservative approach did not allow for any discrepancy between the relative positions in each build. Runs of consecutive genes between which there was no discrepancy in relative positions were used to identify regions in which the position of markers and the subsequent calculation of recombination rates were likely to be

accurate. Recombination windows within such regions were retained. Although relaxation of the size of the discrepancy allowed did not qualitatively affect the results, the most conservative data set was used for all subsequent analyses. Autosomal and X-linked genes were assigned positions based on the midpoint between the start and end of their coding sequence in build 3.1 and, where data was available, these positions were used to assign orthologs a sex-averaged recombination rate in rat. Data were analysed in non-overlapping 1Mb windows.

For both autosomal and X-linked genes, a linear regression weighted by alignment length was performed on recombination rate as a predictor of substitution rate. Comparison of the higher steepness of the autosomal regression compared with that of the X was tested for significance using a one-sided *t*-test,

$$t = \frac{b_x - b_{Autosome}}{\sqrt{\left(s_X^2 + s_{Autosome}^2\right)}}, \tag{8}$$

for which degrees of freedom (*df*) were estimated using the Welch-Satterthwaite equation (Welch, 1947, Satterthwaite, 1946),

$$df = \frac{\left(s_X^2 + s_{Autosome}^2\right)^2}{\left(\frac{s_X^4}{\left(n_X - 1\right)} + \frac{s_{Autosome}^4}{\left(n_{Autosome} - 1\right)}\right)}, \tag{9}$$

where b_N = slope of the regression, s_N = standard error of the mean (SEM) and n_N = sample size of the chromosomal class *N*.

2.2.10 Regionality of substitution rates

The substitution rate of individual Y-linked introns was estimated and subjected to an analysis of variance (ANOVA) by gene. For each autosomal and X-linked gene, the neighbouring 5' and 3' orthologs were identified and the mean of their substitution rates determined. For these chromosomal classes, a Spearman's rank correlation of a given focal gene's substitution rate with the mean of its neighbouring orthologs was performed. A higher steepness of the autosomal regression of focal versus flanking substitution rates compared with that of the X regression was tested for significance using the one-sided *t*-test described previously (Equations 8 and 9, section 2.2.9).

2.2.11 Rearrangement Index

Each mouse autosome was assigned a rearrangement index, a measure of the probability that the rat orthologs of any two randomly selected genes on a given mouse autosome are not both located on the same rat autosome. For a focal mouse autosome, two genes were randomly sampled and the location of their rat orthologs determined. From 1000 samplings, the number of occasions on which the rat orthologs were located on different chromosomes was counted (*n*). The index of rearrangement (RI) was then calculated for the autosome by division of this count by the number of repeat samplings (*n*/1000), such that highly rearranged autosomes were assigned higher indices. Note that this rearrangement index did not quantify the extent of intra-chromosomal rearrangements such as inversions. A linear regression of this index as a predictor of autosomal K_i was then calculated.

2.2.12 Intronic GC content

As previously described for analyses at the genic level, the same indel free intronic alignments that were used to estimate K_i were concatenated by gene. For each genic alignment, counts of each base (A, T, C and G) were made from the two aligned sequences. The sum of G + C bases from the alignment was then divided by the sum of the two sequence lengths, such that the intronic G+C content (GC) was calculated as [(G + C) / (A + T + C + G)].

2.2.13 Calculation of G+T skew – a proxy for germ-line expression rate

For each intronic alignment concatenated by gene, the proportions of A, T, C and G bases in the mouse and the rat sequence were determined. From these base compositions, the extent of G+T bias was calculated as the ratio of [(G+T)-(A+C)]/(G+C+A+T) for each species, from which the mean G+T bias of the two species was calculated. This was used as a proxy for germ-line gene expression in all subsequent analyses.

2.2.14 Control for a different past history of inversions

From a visual inspection of whole genome rodent synteny maps, obtained from MGI (Eppig *et al*., 2007) and Nilsson *et al*. (2001), putatively collinear regions were identified as mouse-rat chromosome pairs 2-3, 3-2, 4-5, 7-1, 9-8, 12-6, 16-11 and 18-

18. Within these regions, inversions were identified from a reversal in the order of orthologous genes on one chromosome. Similarly, single gene rearrangements were identified from breaks of gene order on either chromosome. Exclusion of orthologous genes in either category restricted the autosomal data set to 1558 genes located in collinear regions that had not been subject to intra-chromosomal rearrangements.

2.3 Results

Two data sets were generated: aligned introns purged of those in which conserved residues were clustered (possibly owing to hidden exons) and synonymous rates in exons. As it is known that exonic synonymous mutations can be subject to selection in mammals (Chamary *et al*., 2006), the main analyses presented here focus on what was probably the safest data set, namely, the filtered introns (Figure 2.3). However, the main findings were robust to the use of K_S, K_4, alternative K estimators and alternative measures of centrality, results for which are also given.

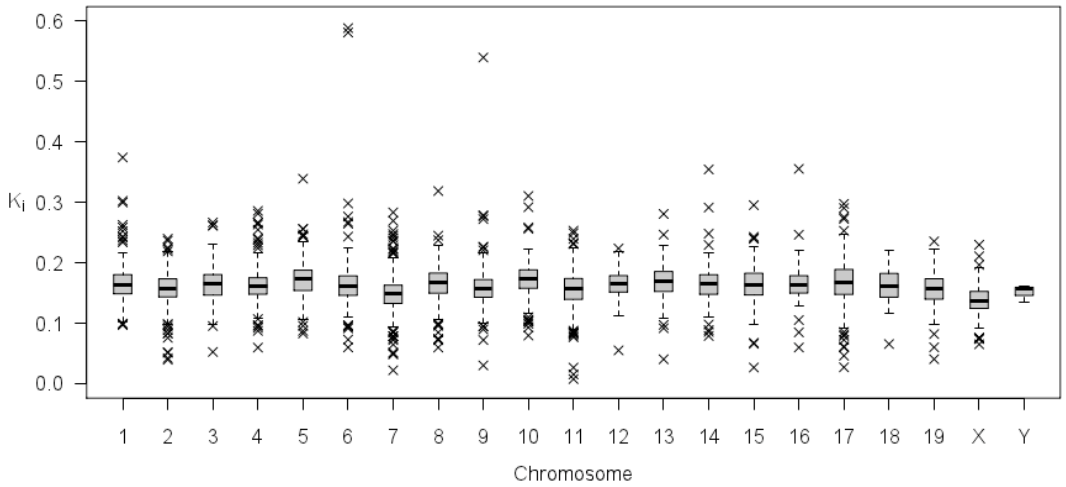


Figure 2.3: Boxplot showing the distribution of genic substitution rates, measured at intronic sites (K_i), across each mouse chromosome, using the filtered data set.

2.3.1 Estimates of α are dependent on the chromosomes used

For the focal intronic data set it was found that rates of evolution were in the order $K_{Autosome}$ = 0.1645 (0.1642, 0.1647) > K_Y = 0.1494 (0.1393, 0.1598) > K_X = 0.1385 (0.1373, 0.1397), with the autosomal rates significantly higher than the Y chromosome (P = 0.0031, Figure 2.4, for statistical test, see Methods section 2.2.8).

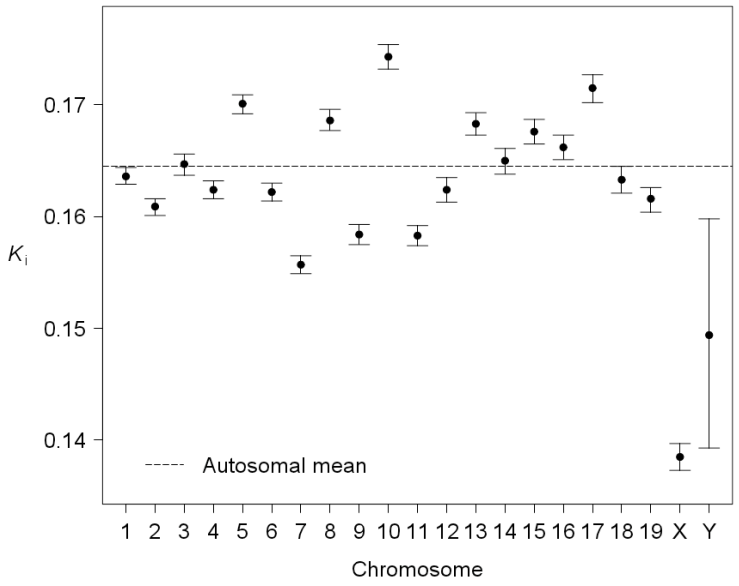


Figure 2.4: Rates of intronic evolution on the X chromosome, Y chromosome and each mouse autosome. For each chromosome, K_i estimated from concatenated intronic sequences and the 95% confidence intervals are shown. The dashed line is the intronic autosomal mean rate of evolution.

The finding that $K_{Autosome} > K_Y > K_X$ was robust to alternative K_i estimators and measures of autosomal centrality (Table 2.2). The same was true at synonymous sites for both K_S and alternative K_4 estimators using alternative measures of centrality. However, due to the relatively small sample size of the Y-linked data set giving rise to wide confidence intervals, differences at synonymous sites were not significant (Table 2.3).

Chromosome	K_i JC	K_i KM	K_i TN	K_i TK
A ($\bar{x}$)	0.1620 (0.1618, 0.1622)	0.1642 (0.1640, 0.1644)	0.1645 (0.1642, 0.1647)	0.1645 (0.1642, 0.1647)
A ($\bar{x}_w$)	0.1614 (0.1612, 0.1616)	0.1636 (0.1634, 0.1638)	0.1639 (0.1637, 0.1641)	0.1639 (0.1637, 0.1641)
A (M)	0.1611 (0.1606, 0.1619)	0.1633 (0.1627, 0.1641)	0.1637 (0.1631, 0.1645)	0.1636 (0.1631, 0.1645)
X	0.1367 (0.1355, 0.1379)	0.1381 (0.1369, 0.1393)	0.1385 (0.1373, 0.1397)	0.1385 (0.1373, 0.1397)
Y	0.1473 (0.1375, 0.1575)	0.1487 (0.1387, 0.1590)	0.1494 (0.1393, 0.1599)	0.1494 (0.1393, 0.1598)

Table 2.2: Chromosomal rates of evolution with 95% confidence intervals derived from intronic sites using alternative *K* estimators where JC = Jukes and Cantor; KM = Kimura; TN = Tamura and Nei and TK = Tamura and Kumar. Alternative measures of autosomal (A) centrality are also given where $\bar{x}$ = mean; $\bar{x}_w$ = weighted mean; and M = median.

K estimator	Measure of Centrality	Autosomal	X-Linked	Y-Linked
K_S	$\bar{x}$	0.1744 (0.1728, 0.1760)	0.1500 (0.1396, 0.1610)	0.1577 (0.1302, 0.1879)
	$\bar{x}_w$	0.1757 (0.1741, 0.1773)	0.1454 (0.1363, 0.1550)	0.1515 (0.1268, 0.1805)
	M	0.1722 (0.1702, 0.1738)	0.1340 (0.1277, 0.1424)	0.1401 (0.1328, 0.2074)
K_4 JC	$\bar{x}$	0.1713 (0.1697, 0.1730)	0.1474 (0.1378, 0.1575)	0.1611 (0.1333, 0.1910)
	$\bar{x}_w$	0.1736 (0.1719, 0.1752)	0.1438 (0.1348, 0.1532)	0.1566 (0.1296, 0.1863)
	M	0.1693 (0.1674, 0.1711)	0.1418 (0.1286, 0.1473)	0.1502 (0.1255, 0.2114)
K_4 KM	$\bar{x}$	0.1741 (0.1724, 0.1759)	0.1497 (0.1398, 0.1601)	0.1625 (0.1337, 0.1928)
	$\bar{x}_w$	0.1764 (0.1747, 0.1781)	0.1459 (0.1368, 0.1550)	0.1580 (0.1313, 0.1873)
	M	0.1719 (0.1699, 0.1736)	0.1435 (0.1289, 0.1494)	0.1517 (0.1256, 0.2138)
K_4 TN	$\bar{x}$	0.1776 (0.1759, 0.1794)	0.1532 (0.1421, 0.1655)	0.1668 (0.1376, 0.1984)
	$\bar{x}_w$	0.1793 (0.1776, 0.1811)	0.1481 (0.1388, 0.1577)	0.1611 (0.1348, 0.1925)
	M	0.1743 (0.1728, 0.1762)	0.1441 (0.1307, 0.1526)	0.1568 (0.1325, 0.2230)

Table 2.3: Substitution rates with 95% confidence intervals at synonymous (K_S) and four-fold degenerate sites (K_4) calculated for each chromosomal class using alternative *K* estimators where JC = Jukes and Cantor; KM = Kimura; and TN = Tamura and Nei. Alternative measures of centrality are also given where $\bar{x}$ = mean; $\bar{x}_w$ = weighted mean; and M = median.

As $K_{Autosome} > K_Y$, it was no surprise that the three comparators failed to agree on the estimate of α, with one supporting a moderate male bias, one a female bias and one no or weak male bias (Figure 2.5) These estimates were not mutually compatible ($P < 0.0001$; for statistical test, see Methods section 2.2.8).

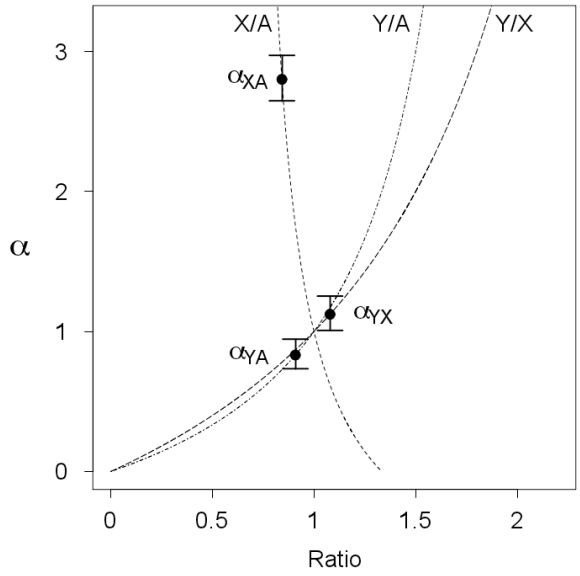


Figure 2.5: Estimates of α from three pairwise chromosomal comparisons under the germ-line replication model. The lines represent the form of the curve relating to the ratio of rates to α for each chromosomal comparison. The 95% confidence intervals were determined from 10,000 bootstraps.

Qualitatively, estimates of α derived from each of the three pairwise between-chromosome comparisons were unaffected by which of the intronic data sets were used (Table 2.4). Similarly, a discrepancy in at least two estimates of α (Table 2.5) was supported by all synonymous substitution rates, despite the potential for selective constraints to act on such sites. That alternative methods of centrality did not affect estimation of α was perhaps unsurprising given that, at least for the autosomal and X-linked datasets, the distributions of each dataset were similar (Figure 2.6). As such, the relative values of each average, on which α is dependent, would remain broadly consistent.

K_i estimator	Measure of autosomal centrality	$\alpha_{XAutosome}$	$\alpha_{YAutosome}$	α_{YX}
K_i JC	$\bar{x}$	2.7596 (2.6088, 2.9202)	0.8339 (0.7375, 0.9459)	1.1210 (1.0077, 1.2481)
	$\bar{x}_w$	2.6998 (2.5526, 2.8550)	0.8391 (0.7417, 0.9523)	1.1210 (1.0077, 1.2481)
	M	2.6655 (2.5197, 2.8478)	0.8422 (0.7428, 0.9551)	1.1210 (1.0077, 1.2481)
K_i KM	$\bar{x}$	2.8213 (2.6684, 2.9955)	0.8276 (0.7311, 0.9393)	1.1197 (1.0054, 1.2485)
	$\bar{x}_w$	2.7589 (2.6101, 2.9255)	0.8329 (0.7354, 0.9454)	1.1197 (1.0054, 1.2485)
	M	2.7241 (2.5724, 2.9140)	0.8359 (0.7371, 0.9454)	1.1197 (1.0054, 1.2485)
K_i TN	$\bar{x}$	2.8017 (2.6482, 2.9731)	0.8320 (0.7343, 0.9453)	1.1229 (1.0076, 1.2530)
	$\bar{x}_w$	2.7394 (2.5900, 2.9037)	0.8373 (0.7387, 0.9518)	1.1229 (1.0076, 1.2530)
	M	2.7162 (2.5638, 2.9071)	0.8393 (0.7395, 0.9535)	1.1229 (1.0076, 1.2530)
K_i TK	$\bar{x}$	2.8011 (2.6480, 2.9724)	0.8320 (0.7343, 0.9454)	1.1229 (1.0076, 1.2528)
	$\bar{x}_w$	2.7385 (2.5897, 2.9033)	0.8374 (0.7387, 0.9517)	1.1229 (1.0076, 1.2528)
	M	2.7055 (2.5638, 2.9071)	0.8403 (0.7395, 0.9535)	1.1229 (1.0076, 1.2528)

Table 2.4: α and 95% confidence intervals derived from each pairwise comparison of intronic substitution rates using alternative K estimators where JC = Jukes and Cantor; KM = Kimura; TN = Tamura and Nei and TK = Tamura and Kumar. Results using alternative measures of autosomal centrality are also given where $\bar{x}$ = mean; $\bar{x}_w$ = weighted mean; and M = median.

***K* estimator**	**Measure of Centrality**	$\alpha_{\text{XAutosome}}$	$\alpha_{\text{YAutosome}}$	α_{YX}
K_S	$\bar{x}$	2.4503 (1.5979, 3.9784)	0.8251 (0.5963, 1.1617)	1.0793 (0.8040, 1.4667)
	$\bar{x}_w$	3.1418 (2.0818, 5.1512)	0.7581 (0.5646, 1.0588)	1.0643 (0.8034, 1.4505)
	M	4.9792 (3.1594, 8.1053)	0.6858 (0.6220, 1.5363)	1.0699 (0.9212, 2.3134)
K_4 JC	$\bar{x}$	2.4406 (1.6338, 3.8666)	0.8877 (0.6359, 1.2573)	1.1464 (0.8524, 1.5662)
	$\bar{x}_w$	3.1199 (2.0896, 5.1005)	0.8220 (0.5954, 1.1601)	1.1398 (0.8482, 1.5487)
	M	2.9009 (2.2346, 6.3506)	0.7972 (0.5834, 1.6912)	1.0916 (0.8173, 2.0993)
K_4 KM	$\bar{x}$	2.4505 (1.6362, 3.9359)	0.8753 (0.6243, 1.2334)	1.1342 (0.8406, 1.5401)
	$\bar{x}_w$	3.1532 (2.1170, 5.1530)	0.8117 (0.5921, 1.1298)	1.1303 (0.8448, 1.5308)
	M	2.9654 (2.1812, 6.6176)	0.7897 (0.5704, 1.6755)	1.0882 (0.8041, 2.1200)
K_4 TN	$\bar{x}$	2.4054 (1.5110, 4.0145)	0.8849 (0.6315, 1.2612)	1.1393 (0.8377, 1.5701)
	$\bar{x}_w$	3.1910 (2.1340, 5.2522)	0.8155 (0.6038, 1.1575)	1.1380 (0.8545, 1.5641)
	M	3.1587 (2.2077, 7.1395)	0.8179 (0.6084, 1.7955)	1.1383 (0.8590, 2.2640)

Table 2.5: α and 95% confidence intervals derived from each pairwise comparison of synonymous (K_S) and four-fold degenerate sites (K_4) calculated for each chromosomal class using alternative *K* estimators where JC = Jukes and Cantor; KM = Kimura; and TN = Tamura and Nei. Results using alternative measures of autosomal centrality are also given where $\bar{x}$ = mean; $\bar{x}_w$ = weighted mean; and M = median.

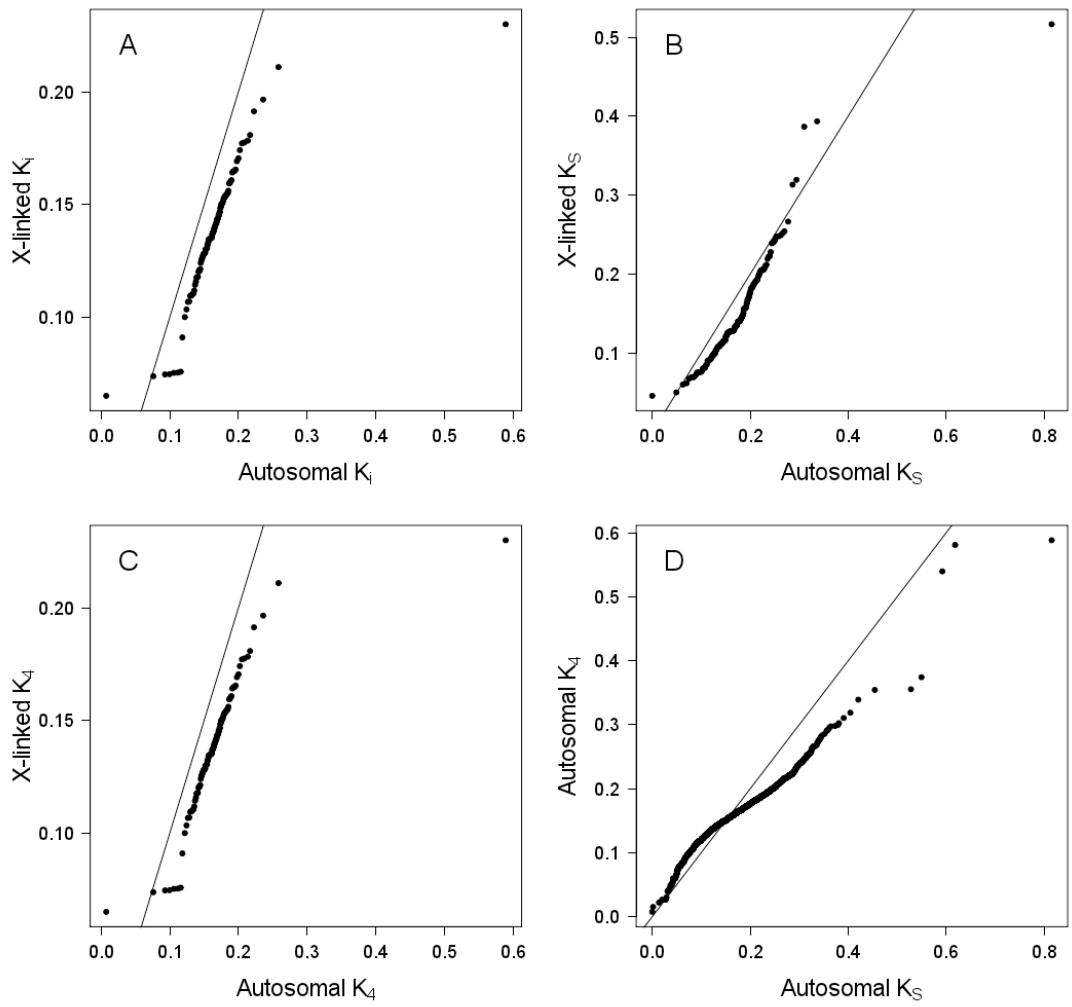


Figure 2.6: Q-Q plots showing comparisons of distributions between autosomal and X-linked data at intronic (A), synonymous (B) and four-fold degenerate (C) sites. Also shown is a comparison of the distributions of the autosomal intronic and synonymous datasets (D). For all plots, solid lines are $y=x$, representing identical distributions.

Why do the three estimates of α not agree? The model of Miyata *et al*. (1987), used to estimate α, assumes that mutational variability is solely determined by variation in the number of DNA replications. However, a number of other potential causes of substitution rate variability have been proposed, including GC content, expression level and differential divergence times due to location within an inversion. The potential for these three parameters to account for the discrepant estimates of α was therefore investigated.

2.3.2 Discrepancies in α are robust to controls for GC content

A positive correlation between GC content and mutation rate has previously been reported in rodents (Hurst and Williams, 2000). As significant differences exist between the GC content of different chromosomal classes (Kruskal-Wallis, P = 1.648 x 10^{-10}) and between different autosomes (Kruskal-Wallis, P < 2.2 x 10^{-16}), might then differences in GC content between the three chromosomal classes have accounted for the observed discrepancy in α?

In order to determine whether GC content could have explained the discrepancy in estimates of α, it was first asked whether variability in K_i across autosomal genes could be attributed to this variation in GC content. A linear regression of GC content as a predictor of K_i for autosomal genes (n = 4051 genes), showed a significantly weak negative relationship (K_i = 0.19759 - 0.07535 GC, r^2 = 0.02303, P = 2.2 x 10^{-16}; Figure 2.7). However, as r^2 was low, this suggested that GC content alone did not determine K_i.

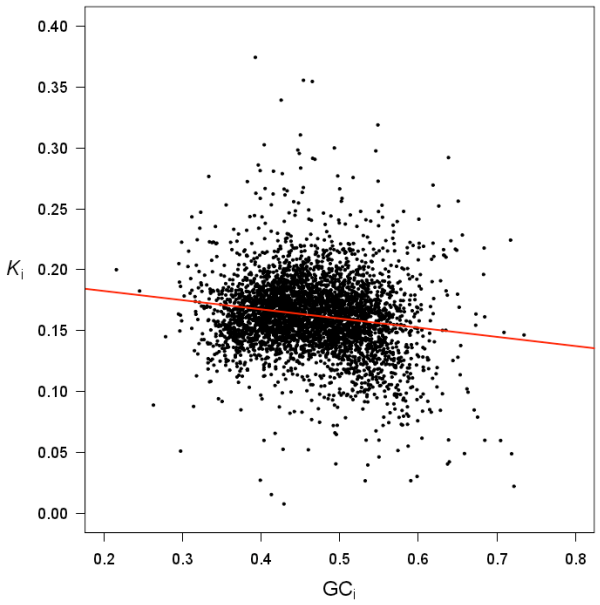


Figure 2.7: K_i declines significantly, albeit weakly, with increasing GC content across 4051 autosomal genes. The line represents the linear least squares regression of K_i against intronic GC content.

Given the differences in both GC and K_i between chromosomes, it was then asked whether chromosome might have had an effect. The assumptions of an analysis of

covariance (ANCOVA) were violated both by a significant interaction ($P = 0.0158$) between GC and autosome and by a different direction of relationship between K_i and GC content on each mouse autosome (Figure 2.8). A Kruskal-Wallis test on the residuals of each autosomal regression of K_i against GC was therefore performed instead. This demonstrated that the amount of residual variation in K_i not explained by GC content differed significantly between each of the autosomes ($P = 3.06 \times 10^{-29}$), suggesting that mouse autosome did have an effect. This was also true when weighting the regression by alignment length ($P = 5.60 \times 10^{-33}$).

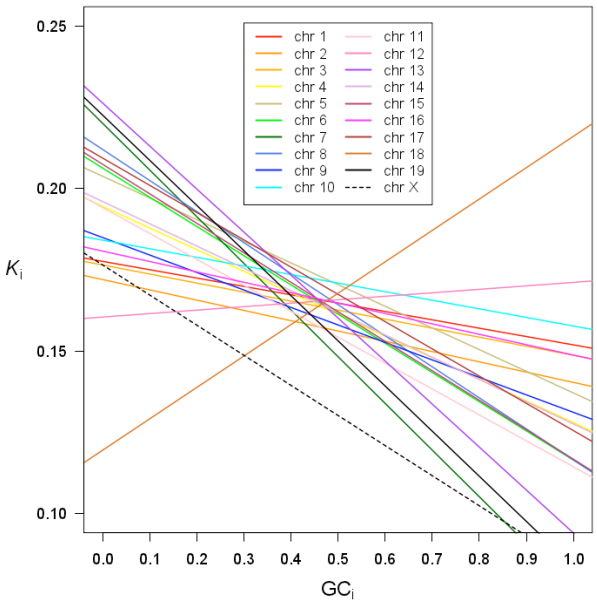


Figure 2.8: The direction of the relationship between GC_i and K_i differs on each of the 19 mouse autosomes. The lines represent linear least-squares regressions of K_i against intronic GC content. Also shown is the line for the X chromosome (dashed).

Nonetheless, to determine whether the relationship between GC content and K_i and the differences in GC content of the three chromosomal classes might account for the observed discrepancies in α, the three pair-wise estimates of α were recalculated, controlling for GC content. To do this a linear least squares regression of autosomal K_i as a response of GC content was used to predict $K_{Autosome}$ from GC_X, the mean GC content of the X-linked data set (Figure 2.9). This predicted $K_{Autosome}$ was then substituted into the X to autosomal comparison to derive $\alpha_{XAutosome}$ (Table 2.6). Similarly, $K_{Autosome}$ was predicted from GC_Y, the mean GC content of the Y-linked data set (Figure 2.9) and substituted into the Y to autosomal comparison to

recalculate $\alpha_{YAutosome}$ (Table 2.6). Finally, from a linear regression of K_i as a response of GC content for the X-linked data set, K_X was predicted from GC_Y (Figure 2.9) and substituted into the Y to X comparison to recalculate α_{YX} (Table 2.6). This process was repeated using a regression weighted by alignment length, using mean GC weighted by alignment length for a given chromosomal class as the predictor of K_i (Table 2.7).

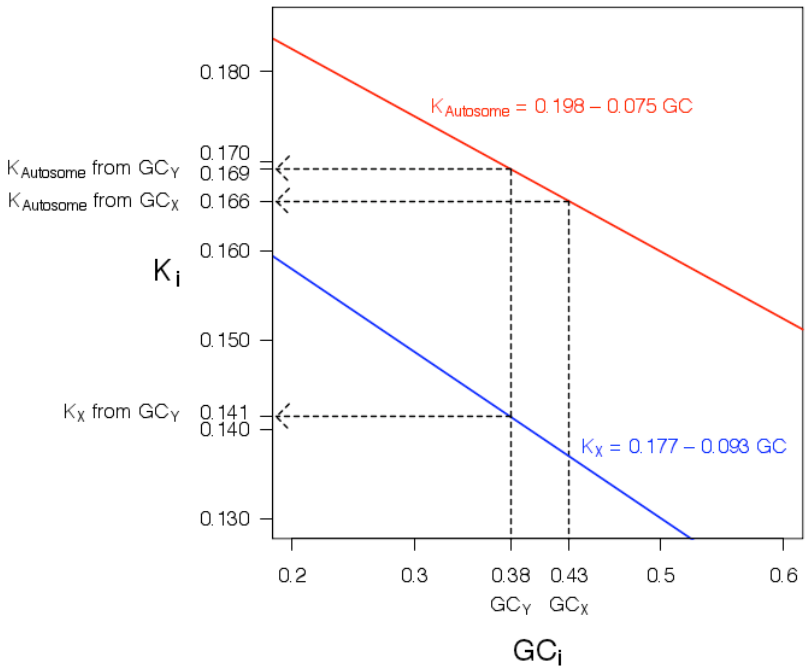


Figure 2.9: Graphical illustration of the method used to control for GC content in the estimate of α, where GC_X is the mean intronic GC content of the X chromosome and GC_Y is the mean intronic GC content of the Y chromosome. Solid lines represent the linear least squares regression of K_i against GC content for the autosomal (red) and X-linked (blue) data sets.

Given the weak negative relationship between K_i and GC content and the lack of consistency of this relationship across the different chromosomes, it was not surprising that controlling for differences in GC content between the three chromosomal classes failed to cause α to converge (Table 2.6). This finding was qualitatively robust to controls for alignment length (Table 2.7).

It could therefore be concluded that although GC content showed a weak relationship with K, it was not sufficient to explain the observed disparity in the three estimates of α.

Comparison	Regression	GC predictor	Predicted K_i	α
X-Autosome	$K_{Autosome} = 0.198 - 0.075$ GC	$GC_X = 0.426$	$K_{Autosome} = 0.166$	3.142
Y-Autosome	$K_{Autosome} = 0.198 - 0.075$ GC	$GC_Y = 0.378$	$K_{Autosome} = 0.169$	0.808
Y-X	$K_X = 0.177 - 0.093$ GC	$GC_Y = 0.378$	$K_X = 0.141$	1.107

Table 2.6: Estimates of α controlling for GC content, derived from a linear regression.

Comparison	Regression	Weighted GC predictor	Predicted K_i	α
X-Autosome	$K_{Autosome} = 0.173 - 0.0189$ GC	$GC_X = 0.3990$	$K_{Autosome} = 0.165$	2.838
Y-Autosome	$K_{Autosome} = 0.173 - 0.0189$ GC	$GC_Y = 0.3779$	$K_{Autosome} = 0.166$	0.826
Y-X	$K_X = 0.1864 - 0.1193$ GC	$GC_Y = 0.3779$	$K_X = 0.141$	1.092

Table 2.7: Estimates of α controlling for GC content and alignment length, derived from a weighted linear regression.

2.3.3 Discrepancies in α are not due to differences in germ-line expression rate

A negative covariance between expression breadth and K_4 has previously been observed (Lercher *et al*., 2004). Unfortunately, the majority of direct measurements of germ-line expression level are derived from terminally differentiated germ cells and as such these may not be representative of expression throughout the germ-line. However, biases in the repair of point mutations associated with transcription coupled repair have been shown to result in an excess of G and T over C and A on the coding strand (Green *et al*., 2003). Importantly, the extent of any bias, indicative of the extent of transcription coupled repair (TCR) has been found to positively correlate with expression intensity in ubiquitously expressed genes - those more likely to be transcribed in the germ-line than tissue specific genes (Majewski, 2003). The extent of G+T bias was therefore used as a proxy for germ-line expression to ask whether variation in germ-line gene expression could account for the between-autosomal variation and the disparity in estimates of α.

No significant differences in the extent of G+T bias were found between the three chromosomal classes (Kruskal-Wallis, $P = 0.8208$) or between different autosomes (Kruskal-Wallis, $P = 0.3020$). Further, a linear regression of G+T bias as a predictor of K_i for autosomal genes ($n = 4051$ genes), showed only a weak negative

relationship that was not significant (K_i = 0.1627 – 0.0040 expression, r^2 = 5.0486 x 10^{-5}, P = 0.6512), suggesting that gene expression did not determine K_i.

Using the same methodology as applied to GC content, estimates of α were recalculated after controlling for differences in G+T bias between the three chromosomal classes. Using both unweighted regressions (Table 2.8) and regressions weighted by alignment length (Table 2.9) to predict either autosomal or X-linked substitution rates as appropriate, the three pairwise estimates of α remained discrepant.

Comparison	Regression	G+T skew predictor	Predicted K_i	α
X-Autosome	$K_{Autosome}$ = 0.163 – 0.004 skew	$Skew_X$ = 0.049	$K_{Autosome}$ = 0.163	2.78
Y-Autosome	$K_{Autosome}$ = 0.163 – 0.004 skew	$Skew_Y$ = 0.071	$K_{Autosome}$ = 0.162	0.871
Y-X	K_X = 0.137 – 0.009 skew	$Skew_Y$ = 0.071	K_X = 0.137	1.166

Table 2.8: Estimates of α controlling for G+T bias, derived from a linear regression. Skew is mean G+T bias for the chromosomal class, a proxy for germ-line expression level.

Comparison	Regression	Weighted G+T skew predictor	Predicted K_i	α
X-Autosome	$K_{Autosome}$ = 0.165 – 0.005 skew	$Skew_X$ = 0.032	$K_{Autosome}$ = 0.165	2.768
Y-Autosome	$K_{Autosome}$ = 0.165 – 0.005 skew	$Skew_Y$ = 0.066	$K_{Autosome}$ = 0.164	0.837
Y-X	K_X = 0.139 – 0.001 skew	$Skew_Y$ = 0.066	K_X = 0.139	1.123

Table 2.9: Estimates of α controlling for G+T bias and alignment length, derived from a weighted linear regression. Skew is mean G+T bias for the chromosomal class, a proxy for germ-line expression.

Given the weak effect of G+T bias on K_i it was unsurprising that controlling for it failed to cause α to converge on a single value. Assuming that G+T skew was a reliable proxy for germ-line expression level, it could therefore be concluded that differential levels of germ-line gene expression between the three chromosomal classes did not account for the observed disparity from the results predicted by the model of Miyata *et al*. (1987).

2.3.4 Controlling for a different past history of inversions does not reconcile α

Do differences between autosomal, X- and Y-linked K_i reflect differences in the mutation rate or in divergence times? Miyata *et al*.'s (1987) model assumes that the

autosomal, X and Y chromosomes diverged at the same time and therefore that differences in divergence at putatively neutral sites stem from differing mutation rates rather than differing times for mutations to accumulate. However, it has been argued that intra-chromosomal genomic rearrangements during speciation, notably inversions, would prevent recombination and as such, cause sequences in rearranged regions to start to diverge earlier (Navarro and Barton, 2003).

Using the sub-sample of 1558 orthologous genes thought not to have been subjected to intra-chromosomal rearrangements (see methods 2.2.14), the mean autosomal substitution rate, weighted by alignment length, was reduced from $K_{Autosome} = 0.1639$ to $K_{Autosome} = 0.1609$ and substituted into Miyata *et al.*'s (1987) equations to estimate α (Table 2.10).

Pairwise comparison	**α**	
	All autosomal genes	**Un-rearranged autosomal genes**
X-Autosome	2.7619	2.3956
Y-Autosome	0.8356	0.8711

Table 2.10: α derived from autosomal comparisons made using the full autosomal data set and after controlling for the effects of a different past history of autosomal inversions.

It was not possible to compile a comparable un-rearranged X-linked data set as all orthologous genes located on the X chromosome were found to have been subject to inversions or single gene rearrangements at some point in their evolutionary history (Figure 2.10). As such, the comparison of a reduced $K_{Autosome}$ with an upper estimate of K_X from the original data set would therefore be expected to give rise to the lowest possible estimate of α_{XA}.

Controlling for autosomal intra-chromosomal rearrangements did not cause α to converge (Table 2.10). To determine whether these results were an artefact of the reduction in autosomal sample size from $n = 4051$ to $n = 1558$, a mantel test was performed. To do this, autosomal genes were randomly sampled, preserving the sample size of the collinear data set. The weighted-mean substitution rate of this sample was substituted into Miyata *et al.*'s (1987) equations for both autosomal

pairwise comparisons, also weighting both mean K_X and K_Y by alignment length. From 10,000 repeat permutations, it was asked how often the pseudo-α was either equivalent to or less than the rearrangement controlled $\alpha_{XAutosome}$ or greater than or equal to the rearrangement controlled $\alpha_{YAutosome}$ as appropriate. Finally, from P = (count + 1) / (number of replications + 1), both $\alpha_{XAutosome}$ and $\alpha_{YAutosome}$ derived from the rearrangement controlled data set were found to be significant (both $P = 9.99 \times 10^{-5}$).

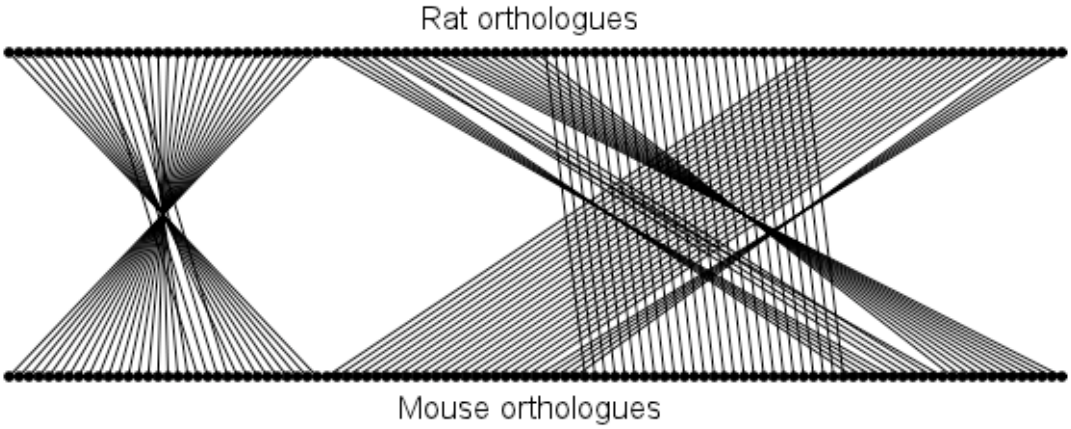


Figure 2.10: Graphical illustration of the relative positions of orthologous genes on the mouse and rat X chromosomes, showing that at least one gene in every pair of orthologs must had been subject to an inversion since divergence from the ancestral X chromosome.

Given that the most conservative method of controlling for a past history of intra-chromosomal rearrangement did not cause α derived from autosomal comparisons to converge and that this was not an artefact of a reduction in sample size, it could be concluded that a disparity in divergence times was unlikely to have accounted for the observed discrepancies in α.

2.3.5 Might recombination also be important?

That autosomes have a higher substitution rate than Y-linked sequence was most unexpected. Why might this have been? Although the replication model has dominated thinking on between-chromosomal class substitution rates, both neutral single nucleotide polymorphism diversity and neutral substitution rates have been found to increase across autosomes in correspondence with the local recombination rate (Lercher and Hurst, 2002, Hellmann *et al*., 2003). This may be owing to

recombination-induced mutation (Perry and Ashworth, 1999, Rattray *et al*., 2001, Lercher and Hurst, 2002, Hellmann *et al*., 2003, Bussell *et al*., 2006) and/or recombination-associated biased gene-conversion (e.g. Dreszer *et al*., 2007 Duret and Arndt, 2008, Berglund *et al*., 2009). A correlation between substitution rate and recombination rate is not, however, universally reported. Both Nachman (2001) and Spencer *et al*. (2006) failed to observe a correlation in humans. That there might be disagreement between studies is unsurprising given that recombination rate data are based on relatively recent crossover events, whereas substitution rates reflect a much longer history. However, it is interesting to note that the pseudoautosomal region of X and Y, a region known to be highly recombining, also exhibits high substitution rates (Perry and Ashworth, 1999, Bussell *et al*., 2006). This region was not included in this analysis.

It was therefore asked whether there was any evidence that in rodents, across autosomes, regions with high recombination rates also had high substitution rates. This issue was, however, enormously problematic. What one needs to know for any sequence is not the current recombination rate alone, but rather the recombination rate to which the sequence has been exposed in both lineages over the course of the divergence of the two species. This is impossible to know. Although it might therefore be better to consider the mean recombination rate of a sequence in mouse and in rat, this too was problematic. The mouse lineage has undergone many rearrangements (Ramsdell *et al*., 2008) so the recombinational environment of a gene in today's mouse genome need not correlate in any manner to its recombinational environment in mouse since the divergence from rat. At the extreme, if a rearrangement was very modern, today's recombinational environment may well be a very poor guide to that which the gene has been exposed over its evolutionary history.

A more defendable test was therefore based on evidence that the rat genome might be vastly more stable than the mouse (Ramsdell *et al*., 2008). Under the assumption that on the megabase scale, each chromosomal region has a characteristic recombination rate (e.g. Paigen *et al*., 2008), then the recombination rate seen in rat might reflect both the recombination rate of a sequence in the rat lineage and that of

some early part of the mouse lineage. It was therefore asked whether rat recombination rates predicted intronic substitution rates. It was found that they do, with a significant relationship between K_i and recombination rate in rat across autosomal genes analysed in non-overlapping 1Mb windows (linear regression weighted by alignment length $r^2 = 0.0346$, $P = 5 \times 10^{-5}$; Figure 2.11).

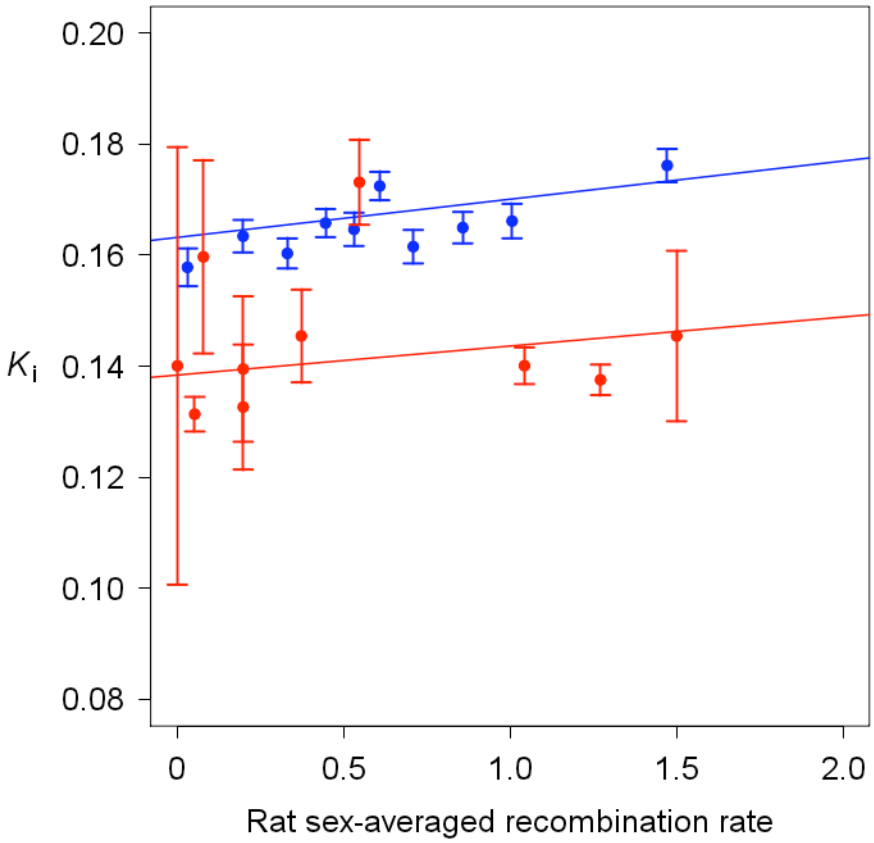


Figure 2.11: The relationship between intronic substitution rate and sex-averaged recombination rate in rat. The points represent averages of bins containing equal numbers of genes, ± standard error of the mean. Autosomal data are in blue, for which bin sample sizes are 111-112 genes. X-linked data are in red, for which bin sample sizes are 2-3 genes. Weighted regression lines are for all data, not the bin means.

The above result would suggest that recombination was positively correlated with substitution rates independent of replication events, all autosomes undergoing the same number of replications. What would be the consequence of this? Such a model could predict that if germ-line replication associated bias, α, were weak (as probably seen in rodents but not necessarily in humans, Makova and Li, 2002), the fact that

recombination in males is limited to autosomes should increase the autosomal substitution rate, possibly exceeding the Y-linked rate.

A simple extension to Miyata *et al.*'s (1987) model was therefore considered, whereby a recombination-associated substitution/mutation effect boosted the rate of evolution by an increment of r. Assuming an equal contribution to the recombination effect from each sex, Equations 2 and 3 were replaced with

$$K_{Autosome} = \frac{\alpha\mu + \mu}{2} + r \qquad (10)$$

and $$K_X = \frac{\alpha\mu + 2\mu}{3} + \frac{2r}{3} \qquad (11)$$

respectively while Equation 1 remained as

$$K_Y = \alpha\mu. \qquad (1)$$

Using data from all three chromosomal classes, these could be solved simultaneously to give α and r:

$$\alpha = \frac{K_Y}{2K_{Autosome} - 3K_X} \qquad (12)$$

$$r = \frac{-4K_{Autosome} + 3K_X + K_Y}{4K_{Autosome} - 6K_X}\mu \qquad (13)$$

From these, it was found that α = 1.7263 (1.5936, 1.8720) and r = 0.5374μ (0.4666μ, 0.6143μ). This suggests that additional replication events in males provide a boost of 0.7263 and recombination supplies a boost of about the same magnitude, probably a little weaker. This finding was robust to the use of alternative K_i estimators (Table 2.11), use of synonymous substitution rates (Table 2.12) and alternative measures of autosomal centrality (Tables 2.11 and 2.12).

K_i estimator	Measure of centrality	α	*r*
K_i JC	$\bar{x}$	1.7095 (1.5788, 1.8499)	0.5250μ (0.4554μ, 0.5989μ)
	$\bar{x}_w$	1.6881 (1.5590, 1.8258)	0.5059μ (0.4373μ, 0.5780μ)
	M	1.6758 (1.5509, 1.8198)	0.4949μ (0.4283μ, 0.5737μ)
K_i KM	$\bar{x}$	1.7304 (1.5978, 1.8768)	0.5454μ (0.4745μ, 0.6228μ)
	$\bar{x}_w$	1.7081 (1.5772, 1.8521)	0.5254μ (0.4558μ, 0.6012μ)
	M	1.6956 (1.5677, 1.8447)	0.5143μ (0.4457μ, 0.5961μ)
K_i TN	$\bar{x}$	1.7265 (1.5939, 1.8721)	0.5376μ (0.4668μ, 0.6144μ)
	$\bar{x}_w$	1.7041 (1.5734, 1.8478)	0.5176μ (0.4479μ, 0.5925μ)
	M	1.6958 (1.5674, 1.8448)	0.5102μ (0.4404μ, 0.5921μ)
K_i TK	$\bar{x}$	1.7263 (1.5936, 1.8720)	0.5374μ (0.4666μ, 0.6143μ)
	$\bar{x}_w$	1.7038 (1.5733, 1.8477)	0.5174μ (0.4479μ, 0.5924μ)
	M	1.6920 (1.5674, 1.8447)	0.5068μ (0.4404μ, 0.5919μ)

Table 2.11: α and *r* with 95% confidence intervals determined from intronic substitution rates, using alternative K_i estimators where JC = Jukes and Cantor; KM = Kimura; TN = Tamura and Nei and TK = Tamura and Kumar. Results using alternative measures of autosomal centrality are also given where $\bar{x}$ = mean; $\bar{x}_w$ = weighted mean; and M = median.

K estimator	Measure of centrality	α	*r*
K_S	$\bar{x}$	1.5598 (1.1146, 2.3340)	0.4452μ (0.1776μ, 0.8959μ)
	$\bar{x}_w$	1.7859 (1.2663, 2.7509)	0.6779μ (0.3404μ, 1.2586μ)
	M	2.4323 (1.6788, 4.6484)	1.2734μ (0.4638μ, 2.2351μ)
K_4 JC	$\bar{x}$	1.6179 (1.1707, 2.3752)	0.4113μ (0.1593μ, 0.8239μ)
	$\bar{x}_w$	1.8586 (1.3221, 2.8377)	0.6306μ (0.3068μ, 1.2072μ)
	M	1.7304 (1.3197, 3.5832)	0.5853μ (0.1888μ, 1.5807μ)
K_4 KM	$\bar{x}$	1.6105 (1.1559, 2.3631)	0.4200μ (0.1690μ, 0.8492μ)
	$\bar{x}_w$	1.8607 (1.3216, 2.8503)	0.6462μ (0.3268μ, 1.2242μ)
	M	1.7497 (1.3157, 3.6466)	0.6078μ (0.1960μ, 1.6524μ)
K_4 TN	$\bar{x}$	1.5987 (1.1124, 2.4468)	0.4033μ (0.1296μ, 0.8646μ)
	$\bar{x}_w$	1.8825 (1.3378, 2.9297)	0.6542μ (0.3242μ, 1.2439μ)
	M	1.8711 (1.3422, 3.9070)	0.6438μ (0.1902μ, 1.7707μ)

Table 2.12: α and *r* with 95% confidence intervals determined from synonymous (K_S) and four-fold degenerate sites (K_4) using alternative *K* estimators where JC = Jukes and Cantor; KM = Kimura; and TN = Tamura and Nei. Results using alternative measures of centrality are also given where $\bar{x}$ = mean; $\bar{x}_w$ = weighted mean; and M = median.

However, this model may not have reflected the whole story. Recent evidence has suggested that the effect of recombination on neutral substitution rates and clusters of biased substitutions correlates strongly with male recombination rates but not with

rates seen in females (Dreszer *et al*., 2007, Berglund *et al*., 2009, Webster *et al*., 2005, Tyekucheva *et al*., 2008, Duret and Arndt, 2008, Galtier *et al*., 2009).

An additional model was therefore considered that excluded female recombination ($r_f = 0$), where Equations 1 and 3 remained unaltered as

$$K_Y = \alpha\mu \tag{1}$$

$$\text{and} \quad K_X = \frac{\alpha\mu + 2\mu}{3}, \tag{3}$$

neither of these chromosomal types recombining in males, but Equation 2 replaced with

$$K_{Autosome} = \frac{\alpha\mu + \mu}{2} + \frac{r_m}{2} \tag{14}$$

where r_m represented a male recombination-associated substitution/mutation effect, to which the autosomes were only exposed to half of the time, whilst in the male germline. As neither the X nor the Y chromosome were subjected to a recombination associated effect, α could therefore be derived directly from Miyata *et al*.'s (1987) Equation 7:

$$\alpha = \frac{\left(2\left(K_Y/K_X\right)\right)}{\left(3-\left(K_Y/K_X\right)\right)} \tag{7}$$

Substitution of Equation 7 into the comparisons of both X to autosome (Equation 3 divided by Equation 14) and Y to autosome (Equation 1 divided by Equation 14), then solving simultaneously, enabled r_m to be derived from:

$$r_m = 2\left(\frac{\left(\left(\frac{2\left(K_Y/K_X\right)}{\left(3-\left(K_Y/K_X\right)\right)}\right)-\left(\frac{\left(K_Y/K_{Autosome}\right)}{2}\right)-\left(\frac{\left(\left(\frac{2\left(Ky/Kx\right)}{3-\left(Ky/Kx\right)}\right)\left(K_Y/K_{Autosome}\right)\right)}{2}\right)\right)}{\left(K_Y/K_{Autosome}\right)}\right)\mu \tag{15}$$

By substituting in data from all three chromosomal classes, it was found that α = 1.1229 (1.0076, 1.2528) and $r_m = 0.3496\mu$ (0.3182μ, 0.3805μ).

Allowance for a possible male only recombination effect therefore suggested a much lower replication-associated bias to the substitution rate. If recombination in males alone is associated with a substitution bias, then these results suggested that in rodents, at least, the effect of replication may have been much overestimated. Again, this finding was robust to the use of alternative K_i estimators (Table 2.13), the use of synonymous substitution rates (Table 2.14) and alternative measures of autosomal centrality (Tables 2.13 and 2.14).

K_i estimator	Measure of centrality	α	r_m
K_i JC	$\bar{x}$	1.1210 (1.0077, 1.2481)	0.3443μ (0.3129μ, 0.3746μ)
	$\bar{x}_w$	1.1210 (1.0077, 1.2481)	0.3559μ (0.3042μ, 0.3663μ)
	M	1.1210 (1.0077, 1.2481)	0.3311μ (0.2998μ, 0.3646μ)
K_i KM	$\bar{x}$	1.1197 (1.0054, 1.2485)	0.3529μ (0.3218μ, 0.3838μ)
	$\bar{x}_w$	1.1197 (1.0054, 1.2485)	0.3444μ (0.3131μ, 0.3755μ)
	M	1.1197 (1.0054, 1.2485)	0.3396μ (0.3083μ, 0.3735μ)
K_i TN	$\bar{x}$	1.1229 (1.0076, 1.1230)	0.3496μ (0.3183μ, 0.3806μ)
	$\bar{x}_w$	1.1229 (1.0076, 1.1230)	0.3411μ (0.3094μ, 0.3721μ)
	M	1.1229 (1.0076, 1.1230)	0.3378μ (0.3058μ, 0.3719μ)
K_i TK	$\bar{x}$	1.1229 (1.0076, 1.2528)	0.3496μ (0.3182μ, 0.3805μ)
	$\bar{x}_w$	1.1229 (1.0076, 1.2528)	0.3410μ (0.3094μ, 0.3720μ)
	M	1.1229 (1.0076, 1.2528)	0.3363μ (0.3058μ, 0.3718μ)

Table 2.13: α and r_m with 95% confidence intervals determined from intronic substitution rates using alternative K_i estimators where JC = Jukes and Cantor; KM = Kimura; TN = Tamura and Nei and TK = Tamura and Kumar. Results using alternative measures of autosomal centrality are also given where $\bar{x}$ = mean; $\bar{x}_w$ = weighted mean; and M = median.

K estimator	Measure of centrality	α	r_m
K_S	$\bar{x}$	1.0793 (0.8040, 1.4667)	0.3081μ (0.1508μ, 0.4726μ)
	$\bar{x}_w$	1.0643 (0.8034, 1.4505)	0.4040μ (0.2539μ, 0.5573μ)
	M	1.0699 (0.9212, 2.3134)	0.5601μ (0.3169μ, 0.6909μ)
K_4 JC	$\bar{x}$	1.1464 (0.8524, 1.5662)	0.2914μ (0.1374μ, 0.4517μ)
	$\bar{x}_w$	1.1398 (0.8482, 1.5487)	0.3867μ (0.2348μ, 0.5469μ)
	M	1.0916 (0.8173, 2.0993)	0.3692μ (0.1588μ, 0.6125μ)
K_4 KM	$\bar{x}$	1.1342 (0.8406, 1.5401)	0.2958μ (0.1445μ, 0.4592μ)
	$\bar{x}_w$	1.1303 (0.8448, 1.5308)	0.3925μ (0.2463μ, 0.5504μ)
	M	1.0882 (0.8041, 2.1200)	0.3780μ (0.1638μ, 0.6230μ)
K_4 TN	$\bar{x}$	1.1393 (0.8377, 1.5701)	0.2874μ (0.1147μ, 0.4637μ)
	$\bar{x}_w$	1.1380 (0.8545, 1.5641)	0.3955μ (0.2448μ, 0.5543μ)
	M	1.1383 (0.8590, 2.2640)	0.3917μ (0.1598μ, 0.6391μ)

Table 2.14: α and r_m with 95% confidence intervals determined from synonymous (K_S) and four-fold degenerate sites (K_4) using alternative *K* estimators where JC = Jukes and Cantor; KM = Kimura; and TN = Tamura and Nei. Results using alternative measures of centrality are also given where $\bar{x}$ = mean; $\bar{x}_w$ = weighted mean; and M = median.

2.4 Discussion

This chapter presented strong evidence that number of replications is not the unique determinant of substitution rate differences between the three chromosomal classes at putatively neutral sites. Making allowances for the weak effects of differences in GC content (Hurst and Williams, 2000), germ-line expression rate (Lercher *et al*., 2004) or a past history of inversions (Navarro and Barton, 2003) did not alter this conclusion. This is important to know as it suggests that the method of Miyata *et al*. (1987), although commonly employed, is fundamentally incorrect. Whether the method is grossly misleading, however, would depend on many parameters.

An unexpectedly elevated autosomal rate of evolution was observed at all sites considered. To account for the autosomes evolving on average faster than the Y-chromosome, two novel models incorporating a recombination-associated mutation or substitution effect were proposed. Both of these models assumed a single substitution rate across the autosomes. Although, as demonstrated both here and elsewhere (Lercher *et al*., 2001, Malcom *et al*., 2003), the autosomes differ significantly in their putatively neutral substitution rates, this was not necessarily

problematic because 1) there is an average autosomal substitution rate and 2) the two new models were an extension of Miyata *et al*.'s (1987) original model which also assumed a single autosomal rate of evolution. Both models also assumed that any recombination-associated substitution effect would be uniform not just across the autosomes but, assuming an equal contribution from each sex, on the X chromosome too. The calculation of a single value for r or for r_m across the genome might therefore have been invalid, not least because the requirement for at least one chiasma per chromosome arm means that recombination rates differ between chromosomes, with shorter chromosomes tending to have higher recombination rates than longer ones (Kong *et al*., 2002, Jensen-Seaman *et al*., 2004, Shifman *et al*., 2006). As a result of both of these assumptions, neither new model would capture inter-atuosomal variation, this instead being compounded in the error terms for the estimates.

The model could therefore have been improved by considering each autosome separately. Firstly, Miyata *et al*.'s (1987) original equations (Equations 5 to 7) could have been re-applied using 19 different X to autosome comparisons and 19 different Y to autosome comparisons. From this it would have been possible to examine the variation in different estimates of both $\alpha_{XAutosome}$ and $\alpha_{YAutosome}$ that had previously been incorporated into the error terms of a single value, and to ask whether these values were mutually compatible both with themselves and between the three pairwise estimates of α. This method could then have been extended to the novel models (Equations 12, 13, and 15) incorporating a recombination-associated effect to determine 19 different values each for α, r and r_m. Again this would have enabled the variance within each estimate to be determined. Further, it would have been informative to ask whether estimates of r and r_m positively covaried with autosomal recombination rates, as might be expected.

The novel models could then have been extended further by incorporating a chromosome specific recombination effect and then asking whether this significantly reduced variation in estimates of α, r and r_m and reconciled each estimate to a single value. To explore how such a modification might be implemented, it is possible to examine the more general model, which assumed an equal contribution of

recombination from males and females. Note that a similar extension could also be applied to the second model that excluded female recombination. In the general model, the recombination-associated substitution effect, *r*, could be modified by a parameter, *l*, specific for each chromosome, such that Equation 1 for the non-recombining Y chromosome would remain unaltered as

$$K_Y = \alpha\mu \quad (1)$$

but Equations 2 and 3 would be replaced with

$$K_{AutosomeN} = \left(\frac{\alpha\mu + \mu}{2}\right) + r\left(\frac{l_{AutosomeN\_male} + l_{AutosomeN\_female}}{2}\right) \quad (16)$$

and $$K_X = \left(\frac{(\alpha\mu + 2\mu)}{3}\right) + r\left(\frac{2}{3} l_{X\_female}\right) \quad (17)$$

respectively where *l* is a chromosome-specific weighting for the recombination-associated effect *r*. This might be a value such as the recombination rate per base pair of chromosome *N*. Note that for the autosomes, half of the effect of *l* is composed of the male-specific per base pair recombination rate and half is female-specific, whereas for the X chromosome the impact of *l* is based only on recombination in females, to which it is subjected two thirds of the time whilst in the female germline. In theory, these models could be solved simultaneously to determine α and *r* for each autosome, however, implementation of these models is left to future work.

Leaving aside these limitations, the results of the original novel models have a number of implications. Firstly, if the true replication-number effect is very large compared with a potential recombination effect (or whatever causes the disparity), then Miyata *et al*.'s (1987) method is unlikely to greatly mislead. This may well be the case in humans where, *a priori*, if replication is associated with mutation, a male bias should be very pronounced. For example, Makova and Li (2002) estimated $K_Y/K_{Autosome}$ to be 1.68 and estimated α to be 5.25. Assuming first a single recombination effect, $r = 0.5375\mu$, then it is possible to use this estimate with Equations 1 and 10 to estimate α to be 10.89. If only a male-specific recombination effect, $r_m = 0.35\mu$, is considered, then this estimate can be used in conjunction with Equations 1 and 14, to correct α to 7.08. Whether it is legitimate to suppose that any recombination effect is the same in rodent and human is unclear, although there is some evidence to suggest that, if anything, in humans it may be stronger (Clément and Arndt, 2011). However, these corrected estimates of α exceed both the original

and proposed $\alpha = 6$ derived from germ-line anatomy. The reason for this approximate insensitivity is that, if α, the replication-associated bias, is high, the relative impact of male recombination on between-chromosome estimates is reduced. Conversely, the evolutionary rates of rodents may be especially instructive, as any recombination and replication effects are likely to be more balanced.

The models above also suggest that whether any recombination effect is associated with males alone may be very important. Given a lack of understanding of any substitution-recombination correlation on a mechanistic basis, it seems impossible to arbitrate at this time. The observation that gene conservation predominantly occurs in the mitotically dividing spermatogonia (Böhme and Högstrand, 1997) might be important, but the regular finding of increased pseudoautosomal evolutionary rates (Perry and Ashworth, 1999, Filatov and Gerrard, 2003, but see Yi *et al*., 2004) is more obviously consistent with meiotic events.

There is some weak evidence consistent with a male bias to recombination-associated substitution bias, but this is not definitive. If male recombination is the sole or dominant source of within-autosome heterogeneity in substitution rates, then one might have expected to see no or lesser regionality of substitution rates on the X chromosome and on the Y chromosome, these never being subject to recombination in males. Although data remained limited on the Y chromosome as it has too few genes on it, an ANOVA reported no gene effect on substitution rates for Y-linked introns ($P = 0.5628$). For the X chromosome and autosome, the rate of evolution of a gene with its immediate chromosomal neighbours (one 5' and one 3') could be compared. On the X chromosome, there was no correlation (Spearman's $\rho^2 = 0.007$, $P = 0.40$), whereas on autosomes, there was a correlation an order of magnitude higher (Spearman's $\rho^2 = 0.054$, $P = 2.2 \times 10^{-16}$; Figure 2.12). As expected, the slope of the regression line of focal versus flanking for autosomes was steeper than that on the X (slope for autosomes = 0.167 ± 0.01 SEM, for X = 0.0472 ± 0.07 SEM, $t = 1.69$, $df = 107$, $P < 0.05$). These data were consistent with a dominant effect of recombination in the male germ-line. However, this test suffered from three problems. First, the gene density on the X chromosome was lower than on the autosomes, so immediate neighbours on the X from the ortholog sample were less

likely to be in the same recombination block. Second, recombination in females is more scattered along chromosomes than in males (Paigen *et al.*, 2008), hence any female effect on the X need not be visible in a comparison between neighbours, while nonetheless a potent force in determining the overall putatively neutral rate of evolution of the X. Finally, the number of genes on the X chromosome was considerably lower than the autosomal samples size and thus the power of each test was not comparable.

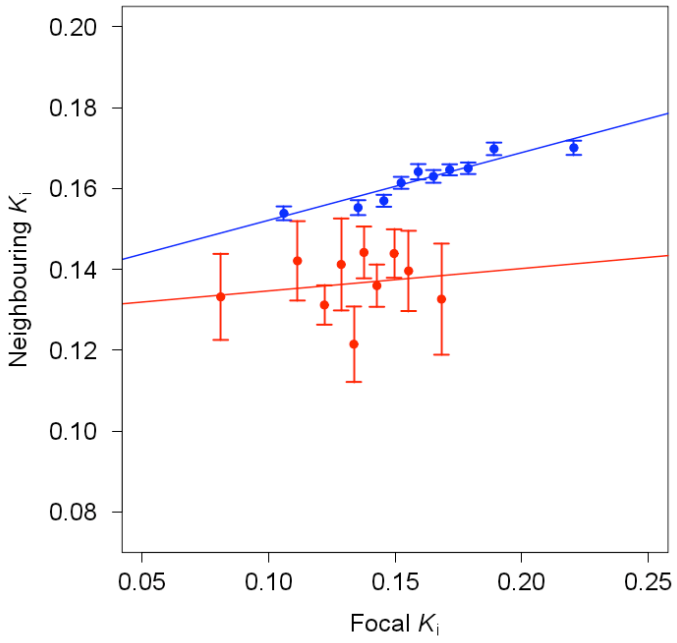


Figure 2.12: Comparison of a focal gene's intronic substitution rate with the mean of its 5' and 3' nearest neighbours. Data shown are bin averages (± SEM) where, for each chromosomal class, bins contain equal numbers of genes, 1000 for the autosomes (blue) and 401 for the X chromosome (red). Regression lines are for all data, not bin means.

Further, if only male recombination is mutagenic, then one would not have expected to see a relationship between recombination rate and substitution rate for X-linked genes, as these do not recombine in males. Indeed, although it was found that recombination rate in rat can predict the intronic substitution rate on the autosomes, no such effect was observed on the X chromosome (weighted linear regression for autosomes $r^2 = 0.0346$, $P = 5 \times 10^{-5}$; for X $r^2 = 0.004$, $P = 0.8122$; Figure 2.11). However, given the weakness of the effect, it was unsurprising that a steeper slope on the autosomes than on the X was not found (slope for autosomes = 0.0086 ±

0.0021 SEM, for X = -0.0026 ± 0.0111 SEM, $t = 1.0$, $df = 13.949$, $P = 0.167$). A female recombination effect could not therefore be completely excluded.

Two further observations could be made. Any theory to explain why the X, Y and autosomes evolve at different rates should also attempt to account for why different autosomes evolve at different rates. The suggested recombination model might be able to explain one curious observation. A striking correlation ($\rho = 0.7488$, $P < 0.00009$) was found between the probability that two randomly chosen genes on a given mouse chromosome have their orthologs on the same chromosome in rat and the evolutionary rate of the mouse chromosome (Figure 2.13). Two factors might link this observation to recombination. First, one must assume that regions associated with high recombination rates have high substitution rates. Such high-recombination, fast-evolving domains may be expected to be associated with genomic rearrangements, first because, at least in some species, rearrangements tend to occur in regions of high recombination (Akhunov *et al*., 2003), and second because when chromosomal fusions and translocations occur, they tend to move telomeres rendering them non-telomeric (Dreszer *et al*., 2007). If high rates of telomeric recombination are associated with increased substitution rates, fusions of such regions should have elevated rates of evolution, as recently reported at the fusion point of human chromosome 2 (Dreszer *et al*., 2007).

Second, if recombination in females has little or no effect on substitution rates but male recombination is important then in birds, in which Z chromosomes can recombine in males, Z-W comparisons would be expected to produce estimates of α that are biased upward. Consistent with this it has been noted (Hurst and Ellegren, 1998) that given their life span, the Z-W derived estimates of α are sometimes unusually high, although this in part may be related to extrapair paternity resulting in sexual selection driving increased sperm production and thus more replication events in males (Bartosch-Härlid *et al*., 2003). Comparably, if male recombination is the cause of disparity between estimators of α, assuming nothing else particular about the X chromosome, X-Y comparisons are probably best to estimate α, as male recombination should not influence these predictions. It is probably for this reason that X-Y comparisons are those that in the past have more accurately reflected

presumed differences in germ-line replication ratios (Li *et al*., 2002, Sandstedt and Tucker, 2005, Goetting-Minesky and Makova, 2006), whereas X-autosome comparisons have suggested remarkably high (Smith and Hurst, 1999a), sometimes impossible ($\alpha > \infty$) (McVean and Hurst, 1997) estimates for α.

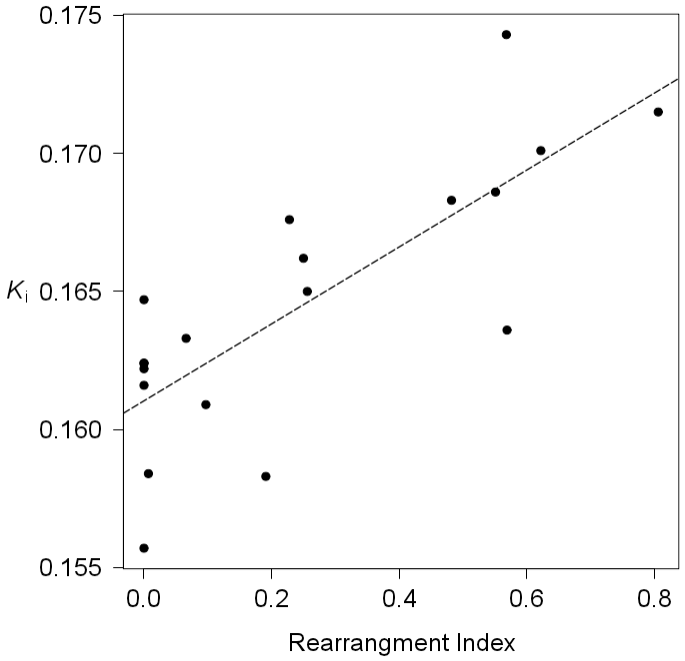


Figure 2.13: The relationship between extent of inter-chromosome rearrangement and rates of intronic evolution of genes on mouse autosomes. As nearly all chromosomal rearrangements occur down the mouse lineage (Ramsdell *et al*., 2008), mouse was employed as the focal chromosome set. Spearman's $\rho = 0.749$, $P = 8.098 \times 10^{-5}$.

2.5 References

AKHUNOV, E. D., GOODYEAR, A. W., GENG, S., QI, L.-L., ECHALIER, B., GILL, B. S., MIFTAHUDIN, GUSTAFSON, J. P., LAZO, G., CHAO, S., ANDERSON, O. D., *et al*. (2003) The organization and rate of evolution of wheat genomes are correlated with recombination rates along chromosome arms. *Genome Research,* 13**,** 753-763.

AXELSSON, E., SMITH, N. G. C., SUNDSTRÖM, H., BERLIN, S. & ELLEGREN, H. (2004) Male-Biased Mutation Rate and Divergence in Autosomal, Z-Linked and W-Linked Introns of Chicken and Turkey. *Mol Biol Evol,* 21**,** 1538-1547.

BACHTROG, D. (2008) Evidence for male-driven evolution in *Drosophila*. *Mol Biol Evol,* 25**,** 617-619.

BARTOSCH-HÄRLID, A., BERLIN, S., SMITH, N. G. C., MØLLER, A. P. & ELLEGREN, H. (2003) Life history and the male mutation bias. *Evolution,* 57**,** 2398-2406.

BAUER, V. L. & AQUADRO, C. F. (1997) Rates of DNA sequence evolution are not sex-biased in *Drosophila melanogaster* and *D. simulans*. *Mol Biol Evol,* 14**,** 1252-1257.

BERGLUND, J., POLLARD, K. S. & WEBSTER, M. T. (2009) Hotspots of biased nucleotide substitutions in human genes. *PLoS Biol,* 7**,** e1000026.

BÖHME, J. & HÖGSTRAND, K. (1997) Timing and effects of template number for gene conversion of major histocompatibility complex genes in the mouse. *Hereditas,* 127**,** 11-18.

BRUDNO, M., DO, C. B., COOPER, G. M., KIM, M. F., DAVYDOV, E., PROGRAM, N. C. S., GREEN, E. D., SIDOW, A. & BATZOGLOU, S. (2003) LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Research,* 13**,** 721-731.

BUSSELL, J. J., PEARSON, N. M., KANDA, R., FILATOV, D. A. & LAHN, B. T. (2006) Human polymorphism and human-chimpanzee divergence in pseudoautosomal region correlate with local recombination rate. *Gene,* 368**,** 94-100.

CASTILLO-DAVIS, C. I., MEKHEDOV, S. L., HARTL, D. L., KOONIN, E. V. & KONDRASHOV, F. A. (2002) Selection for short introns in highly expressed genes. *Nat Genet,* 31**,** 415-418.

CHAMARY, J.-V. & HURST, L. D. (2004) Similar rates but different modes of sequence evolution in introns and at exonic silent sites in rodents: evidence for selectively driven codon usage. *Mol Biol Evol,* 21**,** 1014-1023.

CHAMARY, J. V., PARMLEY, J. L. & HURST, L. D. (2006) Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nat Rev Genet,* 7**,** 98-108.

CHANG, B. H., SHIMMIN, L. C., SHYUE, S. K., HEWETT-EMMETT, D. & LI, W. H. (1994) Weak male-driven molecular evolution in rodents. *Proc Natl Acad Sci USA,* 91**,** 827-831.

CHANG, B. H. J., HEWETT-EMMETT, D. & LI, W. H. (1996) Male-to-female ratios of mutation rate in higher primates estimated from intron sequences. *Zoological Studies,* 35**,** 36-48.

CHOI, S.-K., YOON, S.-R., CALABRESE, P. & ARNHEIM, N. (2008) A germ-line-selective advantage rather than an increased mutation rate can explain some unexpectedly common human disease mutations. *Proc Natl Acad Sci USA,* 105**,** 10143-10148.

CLÉMENT, Y. & ARNDT, P. F. (2011) Substitution patterns are under different influences in primates and rodents. *Genome Biology and Evolution,* 3**,** 236-245.

CONRAD, D. F., KEEBLER, J. E. M., DEPRISTO, M. A., LINDSAY, S. J., ZHANG, Y., CASALS, F., IDAGHDOUR, Y., HARTL, C. L., TORROJA, C., GARIMELLA, K. V., ZILVERSMIT, M., CARTWRIGHT, R., ROULEAU, G. A., DALY, M., STONE, E. A., HURLES, M. E., AWADALLA, P. & PROJECT, G. (2011) Variation in genome-wide mutation rates within and between human families. *Nat Genet,* 43**,** 712-714.

CROW, J. F. (1997a) Molecular evolution - who is in the driver's seat? *Nat Genet,* 17**,** 129-130.

CROW, J. F. (1997b) The high spontaneous mutation rate: is it a health risk? *Proc Natl Acad Sci USA,* 94**,** 8380-8386.
DRESZER, T. R., WALL, G. D., HAUSSLER, D. & POLLARD, K. S. (2007) Biased clustered substitutions in the human genome: the footprints of male-driven biased gene conversion. *Genome Research,* 17**,** 1420-1430.
DRUMMOND, D. A., BLOOM, J. D., ADAMI, C., WILKE, C. O. & ARNOLD, F. H. (2005) Why highly expressed proteins evolve slowly. *Proc Natl Acad Sci USA,* 102**,** 14338-14343.
DURET, L. & ARNDT, P. F. (2008) The impact of recombination on nucleotide substitutions in the human genome. *PLoS Genet,* 4**,** e1000071.
EDGAR, R. C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research,* 32**,** 1792-1797.
ELLEGREN, H. (2007) Characteristics, causes and evolutionary consequences of male-biased mutation. *Proc Biol Sci,* 274**,** 1-10.
EPPIG, J. T., BLAKE, J. A., BULT, C. J., KADIN, J. A., RICHARDSON, J. E. & GROUP, M. G. D. (2007) The mouse genome database (MGD): new features facilitating a model system. *Nucleic Acids Res,* 35**,** D630-D637.
FILATOV, D. A. & GERRARD, D. T. (2003) High mutation rates in human and ape pseudoautosomal genes. *Gene,* 317**,** 67-77.
GALTIER, N., DURET, L., GLÉMIN, S. & RANWEZ, V. (2009) GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates. *Trends Genet,* 25**,** 1-5.
GIBBS, R. A., WEINSTOCK, G. M., METZKER, M. L., MUZNY, D. M., SODERGREN, E. J., SCHERER, S., SCOTT, G., STEFFEN, D., WORLEY, K. C., BURCH, P. E., *et al*. (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature,* 428**,** 493-521.
GOETTING-MINESKY, M. P. & MAKOVA, K. D. (2006) Mammalian male mutation bias: impacts of generation time and regional variation in substitution rates. *Journal of Molecular Evolution,* 63**,** 537-544.
GORIELY, A., MCVEAN, G. A. T., RÖJMYR, M., INGEMARSSON, B. & WILKIE, A. O. M. (2003) Evidence for selective advantage of pathogenic FGFR2 mutations in the male germ line. *Science,* 301**,** 643-646.
GORIELY, A., MCVEAN, G. A. T., VAN PELT, A. M. M., O'ROURKE, A. W., WALL, S. A., DE ROOIJ, D. G. & WILKIE, A. O. M. (2005) Gain-of-function amino acid substitutions drive positive selection of FGFR2 mutations in human spermatogonia. *Proc Natl Acad Sci USA,* 102**,** 6051-6056.
GREEN, P., EWING, B., MILLER, W., THOMAS, P. J., PROGRAM, N. C. S. & GREEN, E. D. (2003) Transcription-associated mutational asymmetry in mammalian evolution. *Nat Genet,* 33**,** 514-517.
HALDANE, J. B. S. (1947) The mutation rate of the gene for hemophilia and its segregation ratios in males and females. *Ann. Eugen.,* 13**,** 262-271.
HELLMANN, I., EBERSBERGER, I., PTAK, S. E., PÄÄBO, S. & PRZEWORSKI, M. (2003) A neutral explanation for the correlation of diversity with recombination rates in humans. *Am J Hum Genet,* 72**,** 1527-1535.
HURST, L. D. (2006) Mutation Rate: Sex Biases. IN KEHRER-SAWATZKI, H. (Ed.) *Encyclopedia of Life Sciences*. John Wiley & Sons.
HURST, L. D. & ELLEGREN, H. (1998) Sex biases in the mutation rate. *Trends Genet,* 14**,** 446-452.

HURST, L. D. & WILLIAMS, E. J. (2000) Covariation of GC content and the silent site substitution rate in rodents: implications for methodology and for the evolution of isochores. *Gene,* 261**,** 107-114.

HUTTLEY, G. A., JAKOBSEN, I. B., WILSON, S. R. & EASTEAL, S. (2000) How important is DNA replication for mutagenesis? *Mol Biol Evol,* 17**,** 929-937.

JENSEN-SEAMAN, M. I., FUREY, T. S., PAYSEUR, B. A., LU, Y., ROSKIN, K. M., CHEN, C.-F., THOMAS, M. A., HAUSSLER, D. & JACOB, H. J. (2004) Comparative recombination rates in the rat, mouse, and human genomes. *Genome Research,* 14**,** 528-538.

JUKES, T. H. & CANTOR, C. R. (1969) Evolution of protein molecules. IN MUNRO, H. N. (Ed.) *Mammalian protein evolution*. New York, Academic.

KAROLCHIK, D., HINRICHS, A. S., FUREY, T. S., ROSKIN, K. M., SUGNET, C. W., HAUSSLER, D. & KENT, W. J. (2004) The UCSC Table Browser data retrieval tool. *Nucleic Acids Res,* 32**,** D493-D496.

KEIGHTLEY, P. D. & GAFFNEY, D. J. (2003) Functional constraints and frequency of deleterious mutations in noncoding DNA of rodents. *Proc Natl Acad Sci USA,* 100**,** 13402-13406.

KIMURA, K. (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide substitution. *Journal of Molecular Evolution,* 16**,** 111-120.

KONG, A., GUDBJARTSSON, D. F., SAINZ, J., JONSDOTTIR, G. M., GUDJONSSON, S. A., RICHARDSSON, B., SIGURDARDOTTIR, S., BARNARD, J., HALLBECK, B., MASSON, G., SHLIEN, A., PALSSON, S. T., FRIGGE, M. L., THORGEIRSSON, T. E., GULCHER, J. R. & STEFANSSON, K. (2002) A high-resolution recombination map of the human genome. *Nat Genet,* 31**,** 241-247.

LERCHER, M. J., CHAMARY, J.-V. & HURST, L. D. (2004) Genomic regionality in rates of evolution is not explained by clustering of genes of comparable expression profile. *Genome Research,* 14**,** 1002-1013.

LERCHER, M. J. & HURST, L. D. (2002) Human SNP variability and mutation rate are higher in regions of high recombination. *Trends Genet,* 18**,** 337-340.

LERCHER, M. J., WILLIAMS, E. J. & HURST, L. D. (2001) Local similarity in evolutionary rates extends over whole chromosomes in human-rodent and mouse-rat comparisons: implications for understanding the mechanistic basis of the male mutation bias. *Mol Biol Evol,* 18**,** 2032-2039.

LI, W.H. (1993) Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *Journal of Molecular Evolution,* 36**,** 96-99.

LI, W.H., YI, S. & MAKOVA, K. (2002) Male-driven evolution. *Curr Opin Genet Dev,* 12**,** 650-656.

MAJEWSKI, J. (2003) Dependence of mutational asymmetry on gene-expression levels in the human genome. *Am J Hum Genet,* 73**,** 688-692.

MAKOVA, K. D. & LI, W.-H. (2002) Strong male-driven evolution of DNA sequences in humans and apes. *Nature,* 416**,** 624-626.

MALCOM, C. M., WYCKOFF, G. J. & LAHN, B. T. (2003) Genic mutation rates in mammals: local similarity, chromosomal heterogeneity, and X-versus-autosome disparity. *Mol Biol Evol,* 20**,** 1633-1641.

MATASSI, G., SHARP, P. M. & GAUTIER, C. (1999) Chromosomal location effects on gene sequence evolution in mammals. *Curr Biol,* 9**,** 786-791.

MCVEAN, G. T. & HURST, L. D. (1997) Evidence for a selectively favourable reduction in the mutation rate of the X chromosome. *Nature,* 386**,** 388-392.

MIYATA, T., HAYASHIDA, H., KUMA, K. & YASUNAGA, T. (1987) Male-driven molecular evolution demonstrated by different rates of silent substitutions between autosome-and sex chromosome-linked genes. *Proceedings of the Japan Academy. Ser. B: Physical and Biological Sciences,* 63**,** 327-331.

NACHMAN, M. W. (2001) Single nucleotide polymorphisms and recombination rate in humans. *Trends Genet,* 17**,** 481-485.

NAVARRO, A. & BARTON, N. H. (2003) Chromosomal speciation and molecular divergence - Accelerated evolution in rearranged chromosomes. *Science,* 300**,** 321-324.

NILSSON, S., HELOU, K., WALENTINSSON, A., SZPIRER, C., NERMAN, O. & STÅHL, F. (2001) Rat-mouse and rat-human comparative maps based on gene homology and high-resolution zoo-FISH. *Genomics,* 74**,** 287-298.

PAIGEN, K., SZATKIEWICZ, J. P., SAWYER, K., LEAHY, N., PARVANOV, E. D., NG, S. H. S., GRABER, J. H., BROMAN, K. W. & PETKOV, P. M. (2008) The recombinational anatomy of a mouse chromosome. *PLoS Genet,* 4**,** e1000119.

PERRY, J. & ASHWORTH, A. (1999) Evolutionary rate of a gene affected by chromosomal position. *Curr Biol,* 9**,** 987-989.

PINK, C. J., SWAMINATHAN, S. K., DUNHAM, I., ROGERS, J., WARD, A. & HURST, L. D. (2009) Evidence that replication-associated mutation alone does not explain between-chromosome differences in substitution rates. *Genome Biology and Evolution,* 2009**,** 13-22.

QIN, J., CALABRESE, P., TIEMANN-BOEGE, I., SHINDE, D. N., YOON, S.-R., GELFAND, D., BAUER, K. & ARNHEIM, N. (2007) The molecular anatomy of spontaneous germline mutations in human testes. *PLoS Biol,* 5**,** e224.

RAMSDELL, C. M., LEWANDOWSKI, A. A., GLENN, J. L. W., VRANA, P. B., O'NEILL, R. J. & DEWEY, M. J. (2008) Comparative genome mapping of the deer mouse (*Peromyscus maniculatus*) reveals greater similarity to rat (*Rattus norvegicus*) than to the lab mouse (*Mus musculus*). *BMC Evol Biol,* 8**,** 65.

RATTRAY, A., MCGILL, C., SHAFER, B. & STRATHERN, J. (2001) Fidelity of mitotic double-strand-break repair in Saccharomyces cerevisiae: a role for SAE2/COM1. *Genetics,* 158**,** 109-122.

SANDSTEDT, S. A. & TUCKER, P. K. (2005) Male-driven evolution in closely related species of the mouse genus Mus. *Journal of Molecular Evolution,* 61**,** 138-144.

SATTERTHWAITE, F. E. (1946) An approximate distribution of estimates of variance components. *Biometrics Bulletin,* 2**,** 110-114.

SHIFMAN, S., BELL, J. T., COPLEY, R. R., TAYLOR, M. S., WILLIAMS, R. W., MOTT, R. & FLINT, J. (2006) A high-resolution single nucleotide polymorphism genetic map of the mouse genome. *PLoS Biol,* 4**,** e395.

SHIMMIN, L. C., CHANG, B. H. & LI, W. H. (1993) Male-driven evolution of DNA sequences. *Nature,* 362**,** 745-747.

SMITH, N. G. & HURST, L. D. (1999a) The causes of synonymous rate variation in the rodent genome. Can substitution rates be used to estimate the sex bias in mutation rate? *Genetics,* 152**,** 661-673.

SMITH, N. G. & HURST, L. D. (1999b) The effect of tandem substitutions on the correlation between synonymous and nonsynonymous rates in rodents. *Genetics,* 153**,** 1395-1402.

SPENCER, C. C. A., DELOUKAS, P., HUNT, S., MULLIKIN, J., MYERS, S., SILVERMAN, B., DONNELLY, P., BENTLEY, D. & MCVEAN, G. (2006) The influence of recombination on human genetic diversity. *PLoS Genet,* 2**,** e148.

TAMURA, K. & KUMAR, S. (2002) Evolutionary distance estimation under heterogeneous substitution pattern among lineages. *Mol Biol Evol,* 19**,** 1727-1736.

TAMURA, K. & NEI, M. (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol,* 10**,** 512-526.

TAYLOR, J., TYEKUCHEVA, S., ZODY, M., CHIAROMONTE, F. & MAKOVA, K. D. (2006) Strong and weak male mutation bias at different sites in the primate genomes: insights from the human-chimpanzee comparison. *Mol Biol Evol,* 23**,** 565-573.

TYEKUCHEVA, S., MAKOVA, K. D., KARRO, J. E., HARDISON, R. C., MILLER, W. & CHIAROMONTE, F. (2008) Human-macaque comparisons illuminate variation in neutral substitution rates. *Genome Biol,* 9**,** R76.

WEBSTER, M. T., SMITH, N. G. C., HULTIN-ROSENBERG, L., ARNDT, P. F. & ELLEGREN, H. (2005) Male-driven biased gene conversion governs the evolution of base composition in human alu repeats. *Mol Biol Evol,* 22**,** 1468-1474.

WELCH, B. L. (1947) The generalization ofstudent's' problem when several different population variances are involved. *Biometrika,* 34**,** 28-35.

YI, S., SUMMERS, T. J., PEARSON, N. M. & LI, W.-H. (2004) Recombination has little effect on the rate of sequence divergence in pseudoautosomal boundary 1 among humans and great apes. *Genome Research,* 14**,** 37-43.

YIN, L., SERI, M., BARONE, V., TOCCO, T., SCARANARI, M. & ROMEO, G. (1996) Prevalence and parental origin of de novo RET mutations in Hirschsprung's disease. *European Journal Of Human Genetics,* 4**,** 356-358.

Chapter 3. **Timing of Replication is a Determinant of Neutral Substitution Rates But Does Not Explain Slow Y Chromosome Evolution in Rodents**

Catherine J. Pink and Laurence D. Hurst

Based on a paper and supplementary information published at:

Molecular Biology and Evolution (2010). 27(5): 1077-1086

3.1 Introduction

Mutation rates, assayed as the substitution rate at putatively neutral sites, are known to vary at different scales across mammalian genomes, but the reasons for this are not well resolved. On the same autosome, genes differ in their synonymous substitution rate (Wolfe *et al*., 1989) with genes of similar substitution rate clustering (Matassi *et al*., 1999, Lercher *et al*., 2001), an effect that is not explained by clustering of genes with similar expression profile (highly/broadly expressed genes tending to have lower substitution rates, Lercher *et al*., 2004). Domains of similarity in substitution rate appear to be defined by synteny blocks, genes within a block having more homogeneity than between blocks (Malcom *et al*., 2003, Webster *et al*., 2004). At a broader level, striking differences have been observed between chromosomes. Not only are there differences between X, Y and autosome (Shimmin *et al*., 1993, Chang *et al*., 1994, Smith and Hurst, 1999, Makova and Li, 2002, Sandstedt and Tucker, 2005, Goetting-Minesky and Makova, 2006, Bachtrog, 2008, Pink *et al*., 2009) but there are also differences between autosomes (Lercher *et al*., 2001, Ebersberger *et al*., 2002, Malcom *et al*., 2003, Gaffney and Keightley, 2005, Pink *et al*., 2009).

Chapter 2 focused on the dominant explanation for differences between X, Y and the average autosomal rate, this being thought to reflect different numbers of cell

division owing to different times spent in male versus female germ-lines (Crow, 1997a, Crow, 1997b, Hurst and Ellegren, 1998, Li *et al*., 2002, Ellegren, 2007). This theory, the theory of male-driven evolution (Miyata *et al*., 1987), assumes that the majority of mutations arise as errors during DNA replication. Mutational variability should therefore reflect only differences in the number of replications sequences undergo. Given that in longer lived species, maintenance of spermatogonia increases the number of germ-line cell divisions in males relative to females, the Y chromosome, which is restricted to males, might be expected to have a higher substitution rate than the autosomes, which are only exposed to additional male germ-line replications half of the time. In turn, autosomes should evolve faster than the X chromosome, which spends only one-third of its time in the male germ-line.

However, as the evidence presented in Chapter 2 showed, in rodents the number of replication events was unable to explain observed differences between chromosomal classes, both in exonic synonymous substitution rates and intronic rates (Pink *et al*., 2009). For both classes of sequence, estimates of the extent of the male bias (α), based on a model presuming that the number of replications is the sole determinant of neutral substitution rates, varied significantly depending on which two chromosomal classes were considered (X vs. autosomes, X vs. Y, Y vs. autosome). Indeed, in strict contradiction of the hypothesis, the autosomes were found to have a similar, if not higher substitution rate than the Y chromosome. The findings in Chapter 2 also confirmed previous reports of considerable between-autosome variability in putatively neutral substitution rates (Matassi *et al*., 1999, Lercher *et al*., 2001, Malcom *et al*., 2003). Neither this observation nor the discrepant estimates of α are consistent with between-chromosomal variability in mutation rates being predominantly determined by the number of germ-line cell divisions. While Chapter 2 proposed a recombination-associated substitution effect as a source of the higher autosomal rate of evolution than that of the non-recombining Y chromosome, the source of the between-autosomal variability in substitution rates remained unresolved. Although a recombination-associated substitution effect could explain some part of the between-autosomal gene variation ($r^2 = 0.035$, $P = 5 \times 10^{-5}$, Chapter 2), it failed to explain any of the between-autosome variation ($r^2 = 0.0103$, $P = 0.68$).

Given that the number of DNA replications could not account for variability in substitution rates between chromosomes or chromosomal classes, what else might have an effect? To account for both the observed inter- and intra-autosomal variability under the replication model, one must suppose different genomic regions are subject to different rates of replication-associated mutations. Indeed, a hitherto unexplored assumption of Miyata *et al.*'s (1987) model is that, per replication, these errors are uniformly distributed throughout the genome. There is however, reason to believe that this might not be the case.

Recent evidence from primates suggests that later replicating regions of the genome have higher rates of neutral divergence and nucleotide diversity than regions replicating earlier (Stamatoyannopoulos *et al.*, 2009). Stamatoyannopoulos *et al.* (2009) postulate that the effect may be owing to a slowing or stalling of replication late in S-phase, possibly owing to a depletion of the deoxynucleotide triphosphate (dNTP) pool or difficulty negotiating heterochromatised templates. They suggest that the slower speed of replication would in turn mean that DNA would be unwound, in a single stranded format, for longer, leaving it more prone to mutation. However, the mechanism is by no means well resolved. Speed of fork progression appears to be a dynamic feature of replication related to other factors such as dNTP availability (Malínsky *et al.*, 2001, Anglana *et al.*, 2003, Koç *et al.*, 2004) and origin density (Conti *et al.*, 2007, Courbet *et al.*, 2008), and it is not yet fully understood how these vary temporally across S-phase. Further, perturbations of relative dNTP concentrations are themselves directly mutagenic (Martomo and Mathews, 2002, Mathews, 2006). Regardless of the mechanistic uncertainties, importantly, replication timing tends to be a relatively fixed property of a genomic domain, remaining stable from cell cycle to cell cycle (Jackson and Pombo, 1998), with GC-rich, gene-rich regions tending to replicate earlier than AT-rich, gene-poor or heterochromatic regions (Woodfine *et al.*, 2004, Karnani *et al.*, 2007, Hiratani *et al.*, 2008). Note, however, the exceptions discussed in the introductory chapter (section 1.3).

This chapter aims to test the validity of the assumption of Miyata *et al.*'s (1987) model, that per replication errors are randomly distributed across the genome. More

specifically, it will address a number of questions: is timing of replication also related to substitution rates in rodents and can it account for the previously observed inter-autosomal variability. Are any differences in replication timing between the three chromosomal classes (X, Y and autosome) causative of differences in substitution rate, previously thought attributable to differences in the number of replications in the two germ-lines and finally, can controls for replication timing resolve the previous discrepancies in the model used to estimate the extent of the male mutation bias? It has only recently become possible to answer these questions due to the novel availability of replication timing data at a 5.8 Kb probe density across all three chromosomal classes in mouse (Hiratani *et al*., 2008), the only mammalian species with such data yet available for the Y chromosome.

3.2 Methods

3.2.1 Calculation of intronic substitution rates

The substitution rate data set curated for the work in Chapter 2 was used for these analyses, for which methodologies are described in detail in sections 2.2.1 to 2.2.6. However, in contrast to Chapter 2, the analyses presented here utilised only intronic substitution rates, for which two alternative data sets were available. The first unfiltered data set comprised all alignments that passed the filters described in sections 2.2.2 and 2.2.3. The second data set was further purged of all introns thought to be evolving under selection, possibly owing to the inclusion of unannotated exons within intronic sequence. This test for clusters of conserved bases, potentially indicative of hidden functional sites was described in section 2.2.4. For both data sets, introns of the same gene were concatenated and the rate of intronic divergence (K_i) was estimated and corrected for multiple hits according to the model of Tamura and Kumar (2002).

3.2.2 Assignment of chromosomal locations

Positions of genes on the mouse genome were defined by the terminal 5' and 3' bp of the coding sequence. These positions were obtained from annotations of the July 2007 assembly (mm9). As mouse replication timing data were assigned genomic coordinates based on the February 2006 assembly (mm8), the stand-alone liftOver

utility and associated chain file mm9ToMm8.over.chain, both obtained from UCSC, were used to convert positions between builds.

3.2.3 Replication timing data

Replication timing data for mouse cell lines prior to differentiation, measured at a 5.8 Kb probe density, were downloaded from http://www.replicationdomiain.org (Weddington *et al*., 2008, Hiratani *et al*., 2008). Positive values were indicative of early replication and negative values were indicative of replication later during S-phase. Four data sets were available. Three comprised replication times derived from embryonic stems cells (ESCs), whereas the fourth set of replication times was derived from induced pluripotent stem cells (iPS). Although the three ESC lines could be regarded as replicate data sets, the same was not necessarily true of the iPS data. Therefore, to justify the inclusion of data derived from iPS cells, for each chromosome, a Spearman's correlation was performed on the raw data for each possible pairwise comparison between the four data sets, enabling a comparison of the strength of correlations within the ESC data to those between any of the ESC data and the iPS data. Correlations in chromosomal replication timing between pairwise ESC lines were no stronger than correlations between any of the ESC lines and the iPS line (Figure 3.1), confirming the finding by Hiratani *et al*. (2008) that replication profiles of iPS cells were indistinguishable from other ESCs. The four cells lines were therefore treated as replicates.

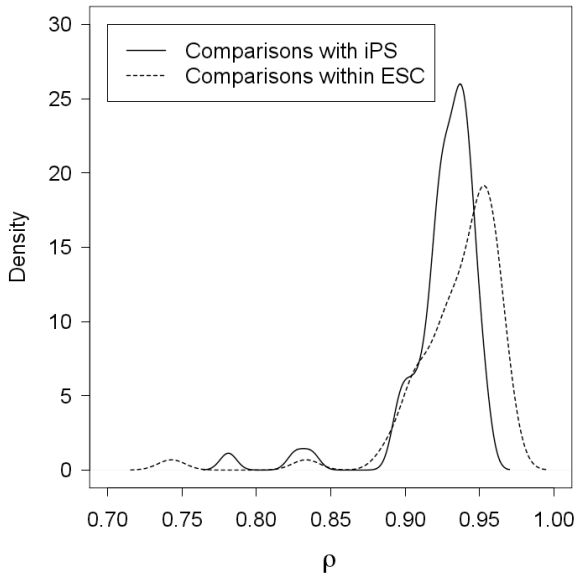


Figure 3.1: Distribution of Spearman's ρ for pairwise correlations of replication timing on each chromosome between embryonic stems cells (ESC) lines (dashed line) and for each ESC with the induced pluripotent stem cells (iPS) (solid line). All spearman's correlations were significant ($P < 0.001$).

3.2.4 Assignment of genic and chromosomal replication times

For each orthologous gene, all replication times obtained from the four cell lines that applied to any part of the gene were identified. This was based on an overlap of the positions of the probe used to calculate replication times and the limits of the coding sequence. A mean of these replication times was then assigned to the gene. From this data set of orthologous genes with both substitution rate and replication time data available, the median intronic substitution rate and median replication time across all genes located on each chromosome were used for analysis at the chromosomal level. 95% confidence intervals were determined from 1,000 bootstraps.

3.2.5 Controls for germ-line expression

Strand asymmetry in the rates of some substitution types has resulted in an excess of G and T over C and A on the coding strand in mammals (Green *et al*., 2003, Mugal *et al*., 2009). This asymmetry is higher in transcribed than in flanking intergenic sequence (Green *et al*., 2003, Mugal *et al*., 2009), and transitions between equal and skewed base composition are clearly associated with the start and end points of transcription (Touchon *et al*., 2003, Polak and Arndt, 2008, Touchon *et al*., 2004).

Together, these observations are strongly suggestive of a germ-line transcription-associated source. Further, the extent of this skew has been found to correlate with expression level in ubiquitously expressed genes (Majewski, 2003). As such genes are more likely to be expressed in the germ-line than tissue-specific genes, the extent of G+T skew was therefore used as a proxy for germ-line expression rate. Unlike the methodology applied in Chapter 2, here G+T skew was only calculated for mouse as it was to be analysed alongside mouse replication timings. For all intronic sequence for a mouse gene, the numbers of A, T, C and G were determined, and the extent of G+T skew was calculated as the ratio $[(G + T) - (A + C)]/(G + C + T + A)$ (Majewski, 2003).

3.2.6 Rearrangement index

Using the method described in Chapter 2 (section 2.2.11), each mouse autosome was assigned a rearrangement index (RI), a measure of the probability that the rat orthologs of any two randomly selected genes on a given mouse autosome are not both located on the same rat autosome. This method was applied to both data sets, namely the final samples of orthologous genes for either the filtered or unfiltered data set as appropriate. Chromosomes having undergone extensive between-autosomal rearrangements were assigned high rearrangement indices, whereas low rearrangement indices were assigned to autosomes that have remained relatively collinear since their common ancestor with rat. Full details of sample sizes, counts and rearrangement indices are supplied in Table 3.1.

Autosome	Number of orthologous mouse genes		Number of times orthologs of sampled mouse pairs were located on different rat autosomes		Rearrangement Index	
	Unfiltered	Filtered	Unfiltered	Filtered	Unfiltered	Filtered
1	331	294	5522	5394	0.5522	0.5394
2	416	356	1095	1029	0.1095	0.1029
3	259	230	0	0	0	0
4	330	288	0	0	0	0
5	351	297	6177	6284	0.6177	0.6284
6	317	272	0	0	0	0
7	384	334	0	0	0	0
8	287	255	5187	5279	0.5187	0.5279
9	315	275	0	0	0	0
10	222	185	5849	5712	0.5849	0.5712
11	412	361	2010	2142	0.201	0.2142
12	164	135	0	0	0	0
13	200	175	4859	4923	0.4859	0.4923
14	189	164	3128	2926	0.3128	0.2926
15	185	156	2342	2392	0.2342	0.2392
16	159	125	2718	2139	0.2718	0.2139
17	214	182	8006	8113	0.8006	0.8113
18	130	112	567	485	0.0567	0.0485
19	195	182	0	0	0	0

Table 3.1: Data used to calculate re-arrangement indices for the unfiltered and filtered data sets.

3.2.7 Calculation of partial spearman correlations

Partial Spearman's correlations between x and y, controlling for z ($\rho_{xy.z}$), were calculated as follows:

$$\rho_{xy.z} = \frac{\rho_{xy} - (\rho_{xz})(\rho_{yz})}{\left(\sqrt{1-\rho_{xz}^2}\right)\left(\sqrt{1-\rho_{yz}^2}\right)} \quad (18)$$

where ρ_{xy} are Spearman's correlations between the two variables indicated by the subscript. Significance was determined by randomly reassigning y to each gene, without replacement, and then re-calculating the partial Spearman's correlation ($\rho_{xy.z}$). This process was repeated 1000 times and the number of occasions (n) on which the strength of the randomised $\rho_{xy.z}$ exceeded that of the original, was used to calculate P as $P = (n + 1) / (1000 + 1)$.

3.3 Results

Two data sets were generated: one subject to a filter for introns thought to contain clusters of sites under selective constraints and a second data set not subject to this filter. As the findings did not, for the most part, qualitatively differ between the two data sets, the main results presented here are from the more conservative, filtered data set. This comprised 4,378 autosomal genes (18.7 Mb), 133 X-linked genes (622 Kb) and 3 Y-linked genes (5.5 Kb). The unfiltered data set comprised 5,060 autosomal genes (79.3 Mb), 150 X-linked genes (2 Mb) and 4 Y-linked genes (20.7 Kb). Where findings qualitatively differed between the two data sets, results for both data sets are given.

3.3.1 Replication time correlates with intronic rates of evolution

First it was asked whether, at the genic level, timing of replication was related to putatively neutral substitution rates. Confirming the previously reported trend seen in primates (Stamatoyannopoulos *et al*., 2009), in rodents there was a significant relationship between timing of replication and intronic substitution rates across both autosomal genes (for the filtered data Spearman's $\rho = -0.0901$, $P = 2.3 \times 10\text{-}9$; Figure 3.2, for the unfiltered data Spearman's $\rho = -0.077$, $P = 3.6 \times 10^{-8}$) and X-linked genes (for the filtered data Spearman's $\rho = -0.2188$, $P = 0.0114$; for the unfiltered data Spearman's $\rho = -0.2417$, $P = 0.0029$). Note that due to the structure of the data, late-replicating sequences were assigned negative timing values, so an increase in any variable during S-phase yielded a negative correlation. For figures, data were plotted on a reversed *x* axis so as to visually show this increase over time. Using the regression $K_i = -0.00440$(replication time) + 0.1717 to predict K_i from the replication times of the first and last genes to replicate, an expected 10.5% increase in rates of evolution during S-phase was observed. However, using the unfiltered data set, for which the regression was $K_i = -0.00335$(replication time) + 0.1742, the magnitude of this increase was reduced to just 7.5%, possibly owing to increased noise in the data set. Both of these values were considerably lower than the 22% increase in divergence reported across primate temporal replication states (Stamatoyannopoulos *et al*., 2009).

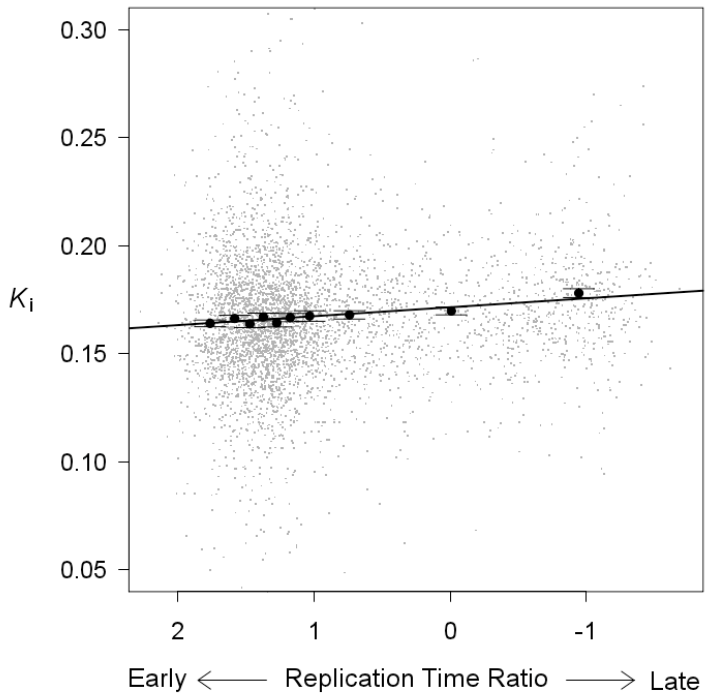


Figure 3.2: Intronic substitution rates increase with later timing of replication across autosomal genes. Spearman's ρ = -0.0901, P = 2.3 x 10^{-9}. Also shown are bin averages (± 1 SEM) for equally sized bins. The linear least squares regression line is for all data, not bin means. Note that the *y*-axis scale results in some outlying data points lying outside the plot area.

3.3.2 GC content did not explain why early replicating genes evolve slowly

Parenthetically, it was interesting to note that the replication time effect ran opposite to a nucleotide-level mutability effect. Consistent with previous work (Woodfine *et al*., 2004, Hiratani *et al*., 2008), a significant, strong correlation between GC content and replication timing across autosomal genes was observed, such that GC-rich sequences replicate early, whereas sequences that are GC-poor replicate late (Spearman's ρ = 0.3153, $P < 2.2 \times 10^{-16}$). GC-rich sequences should therefore evolve slowly owing to replication timing effects. Indeed, a significant, albeit weak, negative relationship (Spearman's ρ = -0.0525, P = 0.00051) was observed. However, it should be noted that this relationship was sensitive to the data set used (for the unfiltered data set Spearman's ρ = -0.0206, P = 0.1437). By contrast, synonymous substitution rates have been found to covary positively with GC content (Hurst and Williams, 2000) and further, CpG dinucleotides are known to be mutable especially when methylated (Coulondre *et al*., 1978).

Given that GC-rich sequences were found to replicate early and be somewhat slow evolving, might this then have accounted for the relationship between replication timing and intronic substitution rates? It was found that it could: when GC content was accounted for the strength of the relationship between replication time and intronic substitution rate was somewhat weaker but remained significant (partial Spearman's $\rho = -0.0777$, $P = 0.0010$), suggesting that this effect was not modulated by GC content. Conversely, most of the relationship between GC and intronic substitution rates was explained by GC-rich sequences being early replicating ($\rho^2 = 0.003$ for uncontrolled analysis, $P = 0.0005$; $\rho^2 = 0.0006$ for the partial correlation controlling for replication timing, $P = 0.042$).

3.3.3 Expression rates do not explain lower rates of evolution of early replicating genes

In mammals, early replication has been associated with gene expression (Holmquist, 1987, Woodfine *et al*., 2004). It might therefore have been the case that the lower substitution rate observed in earlier replicating genes could have been explained by features relating to gene expression. Consistent with this hypothesis a significant correlation between replication time and germ-line expression rate, as assayed by nucleotide skew (Spearman's $\rho = 0.0969$, $P = 1.3 \times 10^{-10}$) was observed, with highly expressed genes replicating earlier. However, it was unclear whether such genes might *a priori* have been expected to have lower rates of evolution, not least because previous evidence has been conflicting. At synonymous sites, the strength of the relationship has ranged from weakly negative (Lercher *et al*., 2004) to non-existent (Duret and Mouchiroud, 2000) and although a significant correlation between intronic rates of evolution and several measures of expression rate has previously been reported in humans (Webster *et al*., 2004), it was not possible to replicate this with the rodent data used here (Spearman's $\rho = 0.0209$, $P = 0.166$). It was therefore unsurprising that using a partial correlation, a significant correlation between replication time and substitution rate remained when controlling for germline expression rate (partial Spearman's $\rho = -0.0926$, $P = 0.0010$). From this it was concluded that the lower substitution rate of earlier replicating genes was not attributable to higher levels of germline gene expression.

3.3.4 Differential timing of replication, in part, explains inter-autosomal variation in substitution rates

The theory of male-driven evolution (Miyata *et al*., 1987) suggests that if the majority of mutations arise as errors during DNA replication, then at the chromosomal level variation in substitution rates should be predominantly determined by differences in the number of replications in each germ-line, more occurring in males than in females of longer lived species. This theory would therefore predict that the autosomes should all evolve at the same rate as, on average, they pass through each germ-line with equal frequencies. However, as discussed in Chapter 2, significant differences in rates of autosomal evolution (Malcom *et al*., 2003, Gaffney and Keightley, 2005, Pink *et al*., 2009) suggest that the number of replications is not the sole determinant of autosomal mutation rates.

Given that genic rates of evolution were found to increase as replication progressed through S-phase, it was therefore asked whether this effect extended to the inter-chromosomal level. First, it was asked whether, on average, the sampled genic sequence located on different autosomes replicated at different times. It was found that they did, with considerable heterogeneity between autosomes in their genic replication timing (Kruskal-Wallis, $P < 2.2 \times 10^{-16}$; Figure 3.3).

It was then asked whether these differences in replication time between autosomes were related to differences in autosomal substitution rates. There was a correlation between replication timing of autosomes and their intronic substitution rate, but whether this was significant depended somewhat on exactly how the data were handled ($\rho_{max} = -0.547$, $\rho_{min} = -0.216$, $P_{min} = 0.017$, $P_{max} = 0.373$; Table 3.2).

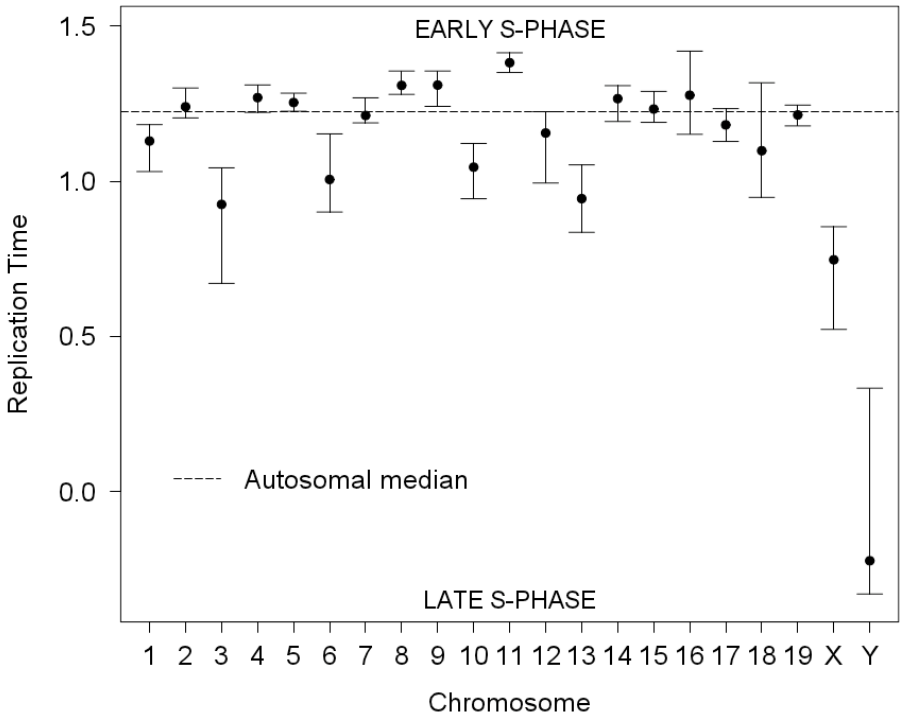


Figure 3.3 Median chromosomal replication times (± 95% confidence intervals) for each of the 19 mouse autosomes and the two sex chromosomes. The horizontal line represents the median across all autosomal genes. There are, on average, significant differences in the timing of replication of different autosomes (Kruskal-Wallis, $P < 2.2 \times 10^{-16}$) and between the three chromosomal classes, X, Y and autosome (Kruskal-Wallis, $P < 2.2 \times 10^{-16}$).

Data set	**Method of calculating autosomal averages**	
	Mean of each autosome	**Median of each autosome**
Unfiltered	Spearman's ρ = -0.3702 P = 0.1194	Spearman's ρ = -0.2158 P = 0.3733
Filtered	Spearman's ρ = -0.5474 P = 0.0168	Spearman's ρ = -0.4002 P = 0.0896

Table 3.2: Spearman's correlations between autosomal replicating timing and intronic substitution rate using alternative data sets and methods of calculating autosomal averages.

Overall, a 4.5% difference in mean rates of intronic evolution between the earliest and latest replicating autosomes was observed, though again this was reduced to only 2.3% when the unfiltered data set was used (based on linear regressions K_i = -0.01603(replication time) + 0.1852 for the filtered data and K_i = -0.00757(replication time) + 0.1791 for the unfiltered data). However, as observed in Chapter 2, for unknown reasons highly rearranged mouse autosomes have high substitution rates

compared to those that have not undergone substantial inter-chromosomal rearrangements (Pink *et al*., 2009). Given the strength of this relationship (r^2 = 0.6063, P = 0.0001; Figure 3.4a), it should be considered alongside any other parameter under investigation as a cause of between-autosome variation in substitution rates, in this instance, timing of replication. Note that extent of inter-autosomal rearrangement is not related to current chromosomal length (Spearman's ρ = -0.045, P = 0.852).

It was therefore asked whether replication timing and amount of inter-chromosomal rearrangement were independent parameters. As no correlation was found between the two variables (Spearman's ρ = 0.0288, P = 0.907) and in a linear model in which both were predictors of autosomal substitution rates, there was no significant interaction term (P = 0.3495, Table 3.3), it was concluded that this was indeed the case.

Call: lm(formula = K_i ~ RT + RI + RT * RI)				
Residuals:				
Min	1Q	Median	3Q	Max
-7.856e-03	-1.089 x 10^{-3}	-1.398 x 10^{-5}	2.073 x 10^{-3}	3.645 x 10^{-3}
Coefficients:				
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.187014	0.009483	19.720	3.87 x 10^{-12}
RT	-0.021051	0.008045	-2.617	0.0194
RI	-0.012124	0.029481	-0.411	0.6867
RT:RI	0.024401	0.025268	0.966	0.3495
Residual standard error: 0.003084 on 15 degrees of freedom **Multiple R-squared:** 0.7474, **Adjusted R-squared:** 0.6968 **F-statistic:** 14.79 on 3 and 15 DF, **p-value:** 9.434 x 10^{-05}				

Table 3.3: Full R output for a linear model in which rearrangement index (RI) and replication timing (RT) were predictors of autosomal intronic substitution rates (K_i). The filtered dataset and autosomal medians were used. Note that the interaction is not significant, P = 0.3495.

Excluding an interaction from the linear model, it was then found that together rearrangement and replication timing could explain a striking 60-70% of between-autosomal variation in substitution rates (for the filtered data r^2 = 0.732, P = 2.689 x 10^{-5}, Table 3.4; for the unfiltered data r^2 = 0.6208, P = 4.274 x 10^{-4}, Table 3.5).

Call: lm(formula = K_i ~ RI + RT)				
Residuals:				
Min	1Q	Median	3Q	Max
-0.0081298	-0.0011638	0.0005467	0.0018111	0.0040737
Coefficients:				
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.180657	0.006812	26.522	1.19×10^{-14}
RT	-0.015578	0.005699	-2.734	0.0147
RI	0.016225	0.002715	5.976	1.94×10^{-5}
Residual standard error: 0.003078 on 16 degrees of freedom **Multiple R-squared:** 0.7317, **Adjusted R-squared:** 0.6981 **F-statistic:** 21.81 on 2 and 16 DF, **p-value:** 2.689×10^{-5}				

Table 3.4: Full R output for a linear model in which rearrangement index (RI) and replication timing (RT) were predictors of autosomal intronic substitution rates (K_i). The filtered dataset and autosomal medians were used.

Call: lm(formula = K_i ~ RT + RI)				
Residuals:				
Min	1Q	Median	3Q	Max
-0.0080829	-0.0021338	0.0007564	0.0020046	0.0062293
Coefficients:				
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.176702	0.007657	23.076	1.04×10^{-13}
RT	-0.008902	0.006450	-1.380	0.186518
RI	0.015768	0.003165	4.982	0.000136
Residual standard error: 0.003566 on 16 degrees of freedom **Multiple R-squared:** 0.6208, **Adjusted R-squared:** 0.5734 **F-statistic:** 13.1 on 2 and 16 DF, **p-value:** 0.0004274				

Table 3.5: Full R output for a linear model in which rearrangement index (RI) and replication timing (RT) were predictors of autosomal intronic substitution rates (K_i). The unfiltered dataset and autosomal medians were used.

Although both parameters contributed significantly to this relationship in the filtered data, rearrangement appeared to be the dominant predictor ($P = 1.94 \times 10^{-5}$ for rearrangement compared with $P = 0.0147$ for replication timing; Table 3.4). This could also be seen by considering how well replication time predicted the residuals of the plot of rearrangement index against autosomal intronic rates (Figure 3.4b). Note however that significance was dependent on the data set used and method used to calculate autosomal centrality (Table 3.6). Similarly, although the significance of replication timing as a co-predictor of autosomal substitution rates in the unfiltered data was sensitive to how autosomal averages were calculated ($P = 0.0299$ for means, Table 3.7; $P = 0.1865$ for medians, Table 3.5), given that significant

relationships were observed in most of the analyses suggested that replication timing should be considered as a covariate in future analyses investigating the causes of variation in autosomal rates of evolution.

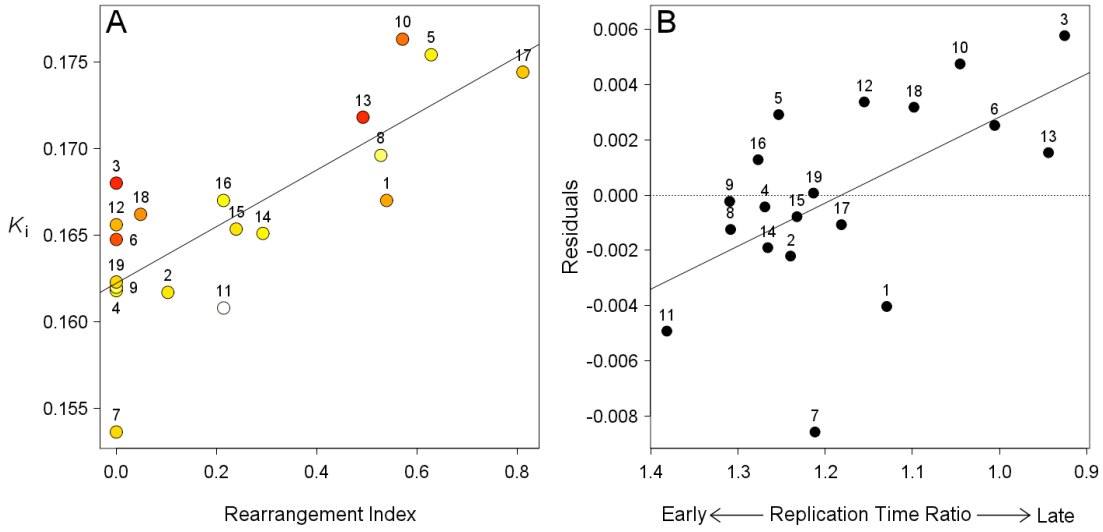


Figure 3.4: (a) The intronic substitution rates of the 19 (labelled) mouse autosomes are significantly predicted by the amount of rearrangement the autosome has undergone (r^2 = 0.606, P = 0.0001, least squares linear regression line shown) and timing of replication (residuals test r^2 = 0.318, P = 0.0119, darker points being indicative of later replication timings). Note the tendency for later replicating autosomes to sit above the line and early replicating ones to sit below. This is further illustrated in (b), a plot of the residuals for (a) against replication time.

Data set	**Measure of autosomal centrality**	
	Means	**Medians**
Filtered	$r^2 = 0.4411$ $P = 0.0019$	$r^2 = 0.3183$ $P = 0.0119$
Unfiltered	$r^2 = 0.2619$ $P = 0.0251$	$r^2 = 0.1062$ $P = 0.1733$

Table 3.6: Residual variation in autosomal K_i that could not explained by extent of rearrangement was subsequently predicted by replication timing using a linear regression. Results of this second regression are given using alternative data sets and measures of autosomal centrality.

Call: lm(formula = K_i ~ RT + RI)				
Residuals:				
Min	1Q	Median	3Q	Max
-0.0069884	-0.0017083	0.0008158	0.0018587	0.0041317
Coefficients:				
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.172652	0.002180	79.216	$< 2 \times 10^{-16}$
RT	-0.005312	0.002229	-2.383	0.0299
RI	0.015180	0.002733	5.554	4.36×10^{-5}
Residual standard error: 0.003081 on 16 degrees of freedom **Multiple R-squared:** 0.6938, **Adjusted R-squared:** 0.6555 **F-statistic:** 18.12 on 2 and 16 DF, **p-value:** 7.732×10^{-5}				

Table 3.7: Full R output for a linear model in which rearrangement index (RI) and replication timing (RT) were predictors of autosomal intronic substitution rates (K_i). The unfiltered dataset and autosomal means were used.

3.3.5 Average replication time of X, Y and autosomal genes were different but controlling for replication time did not account for discrepancies in estimates of α

As shown in Chapter 2, contrary to the predictions of the theory of male-driven evolution, Y-linked introns had a rate of evolution that was at most on a par with those of the autosomes, if not somewhat lower (Pink *et al*., 2009). This was also true considering synonymous sites (McVean and Hurst, 1997, Smith and Hurst, 1999, Pink *et al*., 2009). More generally, estimates of α, the degree of male bias, derived using the method of Miyata *et al*. (1987), were not mutually compatible when using data from the three possible pairwise comparisons (X and autosomes, Y and autosomes and X and Y). Given that later replication timing elevated substitution rates both at the genic and the autosomal level, these discrepancies might have been accounted for if the autosomes replicated later during S-phase than the sex chromosomes. Did then autosomal, X- and Y-linked genes replicate at different times and were autosomal genes on average late replicating compared with those on the Y?

It was found that genes located on each of the three chromosomal classes did replicate, on average, at significantly different times (Kruskal-Wallis test, $P < 2.2 \times 10^{-16}$ for both data sets). Contrary to the above hypothesis however, autosomal genes replicated earliest during S-phase, followed by X-linked genes, with Y-linked genes replicating later in S-phase (median replication times for the filtered data set:

Autosomes = 1.224, X = 0.747, Y = -0.223, Figure 3.3; Median replication times for the unfiltered data set: Autosomes = 1.225, X = 0.762, Y = -0.276.).

It is worth noting that the small sample of Y-linked genes were derived from two BACS and, as such, are positioned close together and therefore subject to similar regional effects, including replication time. It was therefore feasible that the difference in replication time observed across the three chromosomal class samples might have arisen from the Y-linked sample being located in a particularly late-replicating domain. However, the distribution of replication times of the sample genes relative to all probes for a given chromosome (Figure 3.5) showed that this was not the case, with sample genes being clustered in earlier replicating sequences on all chromosomes.

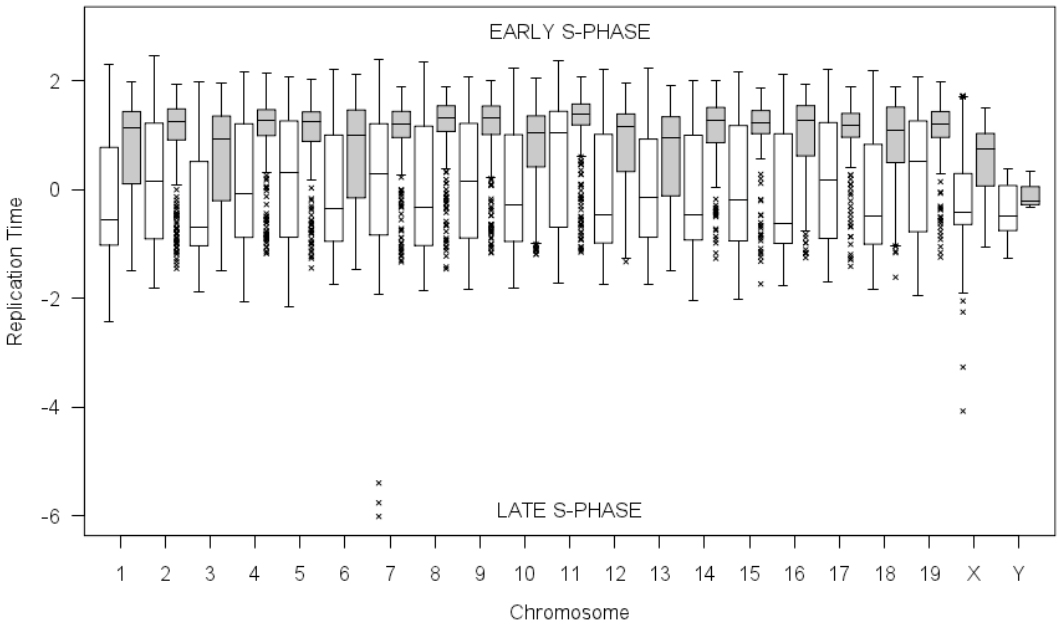


Figure 3.5: Boxplot showing the distribution of replication times across each chromosome. For each chromosome the unfilled box on the left represents replication times across all probes and the grey box on the right represents the replication times of sample genes. For all chromosomes, sample genes are clustered in early replicating sequence.

Given the above result, it was to be expected that the addition of replication time as a covariate would not resolve discrepant estimates of α. Y-linked genes should have a very fast rate of evolution both because they undergo more replication events and because they are relatively late replicating. To understand the quantitative impact of

replication time on estimates of α, a covariate controlled analysis of a form described in Chapter 2, section 2.3.2 was performed.

In order to control for replication time in estimation of the extent of male bias in the mutation rate, a single replication time was imposed across all three chromosomal classes and the magnitude of α calculated using the predicted rate of evolution of each chromosomal class at this time. Because the limited sample size available for the Y chromosome prevented use of a regression of Y-linked genes, the mean replication time across Y-linked genes was used to predict both autosomal and X-linked K_i. This was done using the equation for the least squares linear regression line of replication time as a predictor of K_i across all autosomal and X-linked genes respectively. The ratio of K_Y to the predicted estimate of $K_{Autosome}$ was then inserted into the equation of Miyata *et al*. (1987) to determine $\alpha_{YAutosome}$ from $(K_Y/K_{Autosome})/(2 - (K_Y/K_{Autosome}))$. Similarly, the ratio of K_Y to the predicted K_X was used to calculate α_{YX} from $(2(K_Y/K_X))/(3 - (K_Y/K_X))$. Finally, the predicted estimate of $K_{Autosome}$ relative to the predicted K_X was used to calculate $\alpha_{XAutosome}$ from $3(K_X/K_{Autosome}) - 4)/(2 - 3(K_X/K_{Autosome}))$. As expected, it was found that controlling for replication time failed to reconcile α to a single estimate (Table 3.8).

Class Comparison		Regression	Predicted K_i	α	
				Original	Control for RT
Filtered dataset	X to Autosome	$K_{Autosome} = 0.1717 - 0.0044(RT_Y)$ $K_X = 0.1458 - 0.0113(RT_Y)$	$K_{Autosome} = 0.172$ $K_X = 0.1466$	2.9160	2.5887
	Y to Autosome	$K_{Autosome} = 0.1717 - 0.0044(RT_Y)$	$K_{Autosome} = 0.172$	0.9087	0.8646
	Y to X	$K_X = 0.1458 - 0.0113(RT_Y)$	$K_X = 0.1466$	1.2218	1.1380
Unfiltered dataset	X to Autosome	$K_{Autosome} = 0.1742 - 0.0033(RT_Y)$ $K_X = 0.1532 - 0.0096(RT_Y)$	$K_{Autosome} = 0.1747$ $K_X = 0.1546$	2.3313	2.0489
	Y to Autosome	$K_{Autosome} = 0.1742 - 0.0033(RT_Y)$	$K_{Autosome} = 0.1747$	0.9669	0.9291
	Y to X	$K_X = 0.1532 - 0.0096(RT_Y)$	$K_X = 0.1546$	1.2160	1.1381

Table 3.8: Estimates of α controlling for a single timing of replication, where RT is replication time and K the intronic substitution rate of X, Y or autosomes as indicated by the subscript. For the unfiltered data set, $RT_Y = -0.154$; for the filtered data set, $RT_Y = -0.073$.

3.4 Discussion

Current theory suggests that at the genome-wide level, errors introduced during DNA replication are the primary source of new mutations. An assumption of this theory is that the number of replications is the key determinant of variation in rates of evolution. However, it was shown here that, across autosomal genes where the number of replications is the same, the timing of replication is a significant predictor of rates of evolution. Further, it was shown that replication timing, in conjunction with rearrangement, is a significant predictor of autosomal rates of evolution and that together, these two parameters could explain around 70% of between-autosomal variation in substitution rates.

However, although it was found that on average the sex chromosomes replicate later than the autosomes, they did not exhibit the elevated rates of evolution that might have been expected if a later timing of replication is associated with a higher input of substitutions. In fact, given that the Y chromosome undergoes an increased number of germ-line cell divisions relative to the autosomes and that Y-linked genes replicate, on average, later during S-phase, it might have been expected their rate of evolution was substantially greater than that of autosomal genes. However, this was not found, with the Y-linked genes, if anything, evolving possibly slower than autosomal genes ($K_{\text{Autosome}} = 0.1676 > K_{\text{Y}} = 0.1595$, although significance and magnitude were sensitive to the filters applied to the data set). It was therefore unsurprising that controlling for differences in replication time across the three chromosomal classes failed to cause the three estimates of α to converge. In Chapter 2, it was suggested that an effect of recombination promoting neutral substitutions (either owing to direct mutational effects or owing to biased gene-conversion-like processes) might be an important modulator of substitution rates (Pink *et al*., 2009). That the sampled Y-linked genes were located on the non-recombining Y-specific region of the Y chromosome and that these evolve slower than expected both when considering replication time and number, only further reinforced the paradox.

This result aside, these results had one potential important corollary. At the genic level, replication timing appeared to be an important determinant of substitution

rates in both rodents (as shown here) and in primates (Stamatoyannopoulos *et al*., 2009). If this relationship also holds true in other species, then prior estimates of α that utilise small sample sizes are almost inevitably going to be quantitatively inaccurate. To be more precise, for estimates of α to be inaccurate all that would be needed is that the replication timing of the sequence from one comparator chromosomal class to be different from that of the other. This would be particularly acute for α derived from the X-to-autosomal comparison as this comparison is extremely sensitive to the ratio of rates of evolution (Figure 2.5). Even small inaccuracies in the measurement of substitution rates on either of these chromosomal classes, stemming from a biased sample with respect to position and consequently replication time, would therefore be amplified in inaccurate estimation of α. The problem is potentially even more profound for analyses that compare one Y-linked gene with its X-linked homolog, where, given tiny sample sizes, a major difference in replication timing of the two genes could greatly skew any estimate. If, as shown here, Y-linked sequences generally replicate later than those on the X chromosome, and if this in turn accelerates their evolutionary rate, the finding that Y-linked sequences are fast evolving relative to those on the X chromosome is to be expected, regardless of any differences in the number of replication events. Therefore, the general expectation is that estimates of the magnitude of the male mutation bias derived from an X to Y comparison, employing the equation of Miyata *et al*. (1987), are likely to provide overestimates.

These observations might also explain why the synteny block appears to represent a unit of homogeneity in mutation rate variation (Malcom *et al*., 2003, Webster *et al*., 2004). Replication domains, large regions of similarly timed replication clusters, can range from hundreds of kilobases (Karnani *et al*., 2007) to several megabases (Hiratani *et al*., 2008). In contrast, the scale of mutational variation has been demonstrated to be no larger than 1Mb (Gaffney and Keightley, 2005). It is therefore possible that genomic rearrangements might have moved regions of similarly timed replication with associated similarity in substitution rates, into a genomic landscape differing in replication time and therefore substitution rate (Yaffe *et al*., 2010). An early replicating block of sequence with a slow rate of evolution, for example, could move into a domain of fast evolving sequence or *vice versa*. If the event was

relatively recent, or if the domains brought their replication timing with them as appears to be the case (Yaffe *et al*., 2010), heterogeneity between synteny blocks would result.

The analysis here comes with at least one important caveat. It is possible that the replication timing data used in this analysis did not accurately reflect replication timings in the germ-line. The data used, possibly superior to the somatic data used by Stamatoyannopoulos *et al*. (2009), comprised replication times derived from pluripotent cells as a proxy for germ-line replication times. Given that correlations were observed between timing of replication and other genomic features was suggestive that the data used did reflect germ-line replication times. However, it is known that differentiation is related to temporal changes in replication for as much as 20% of the genome (Hiratani *et al*., 2008). Although the relationship between replication timing and gene expression is not fully understood, it has been suggested that in ESCs, lineage-specific genes may be transcriptionally silent but retain RNA polymerase promoter occupancy and as such replicate early. Upon differentiation, the transcriptional potential of these silent lineage-specific genes is removed and replication occurs later (Azuara *et al*., 2006, Farkash-Amar *et al*., 2008). During the process of gametogenesis, it is therefore likely that some regions of the genome, particularly those containing transcriptionally silent genes, would undergo such shifts in replication time. Such changes would not necessarily be conserved between oogenesis and spermatogenesis nor be distributed uniformly across the three chromosomal classes. Incorrect assignment of germ-line replication times to any of the three chromosomal classes would therefore affect relationships with substitution rates and controls for the estimation of α.

These analyses also suppose that replication time effects have the same mutational effect on X, Y and autosome. Might it be that the sex chromosomes are exposed to a different replication environment during S-phase? Although the formation of the XY body in the male germ-line might represent one such condition, this is unlikely to have an effect because it forms during meiotic prophase, after DNA replication has been completed, and could not therefore differentially influence the effect of replication timing on substitution rates between the chromosomal classes.

Alternatively, such an effect might be female specific, involving X inactivation whereby, on average half of the time, one X chromosome is subject to transient germ-line X inactivation and subsequently replicates late during S-phase. However, this too cannot account for the slow Y-linked rate of evolution, relative to that of the autosomes, because neither of these two chromosomal classes would be affected.

Assuming these caveats to be of minor importance, the results presented here provide evidence in support of replication timing as a source of genomic variation in substitution rates and can potentially explain the previously enigmatic variation in substitution rates between synteny blocks. Although these effects only deepen the mystery of why Y-linked sequence in rodents is not especially fast evolving, more generally it opens up the possibility that all prior calculations of the extent of the male mutation bias, assuming as they do that number of replication events alone is the important determinant, are likely to be wrong. The extent to which prior estimates have misled will depend on the magnitude of the replication timing effect and the difference in timing between the sequences employed. In addition to the possible influence of recombination on differences between X, Y and autosomes, raised in Chapter 2, in the absence of corrective data, these results provide a further reason to strongly caution against the use of Miyata *et al*.'s (1987) equations. Further they argue against the use of single genes or clustered genes in estimation of the impact of the number of germ-line divisions on the mutation rate in male and female germ-lines without adequate control for replication time effects.

3.5 References

ANGLANA, M., APIOU, F., BENSIMON, A. & DEBATISSE, M. (2003) Dynamics of DNA replication in mammalian somatic cells: nucleotide pool modulates origin choice and interorigin spacing. *Cell,* 114**,** 385-394.

AZUARA, V., PERRY, P., SAUER, S., SPIVAKOV, M., JØRGENSEN, H. F., JOHN, R. M., GOUTI, M., CASANOVA, M., WARNES, G., MERKENSCHLAGER, M. & FISHER, A. G. (2006) Chromatin signatures of pluripotent cell lines. *Nat Cell Biol,* 8**,** 532-538.

BACHTROG, D. (2008) Evidence for male-driven evolution in *Drosophila*. *Mol Biol Evol,* 25**,** 617-619.

CHANG, B. H., SHIMMIN, L. C., SHYUE, S. K., HEWETT-EMMETT, D. & LI, W. H. (1994) Weak male-driven molecular evolution in rodents. *Proc Natl Acad Sci USA,* 91**,** 827-831.

CONTI, C., SACCÀ, B., HERRICK, J., LALOU, C., POMMIER, Y. & BENSIMON, A. (2007) Replication fork velocities at adjacent replication origins are coordinately modified during DNA replication in human cells. *Mol Biol Cell,* 18**,** 3059-3067.

COULONDRE, C., MILLER, J. H., FARABAUGH, P. J. & GILBERT, W. (1978) Molecular basis of base substitution hotspots in *Escherichia coli*. *Nature,* 274**,** 775-780.

COURBET, S., GAY, S., ARNOULT, N., WRONKA, G., ANGLANA, M., BRISON, O. & DEBATISSE, M. (2008) Replication fork movement sets chromatin loop size and origin choice in mammalian cells. *Nature,* 455**,** 557-560.

CROW, J. F. (1997a) Molecular evolution - who is in the driver's seat? *Nat Genet,* 17**,** 129-130.

CROW, J. F. (1997b) The high spontaneous mutation rate: is it a health risk? *Proc Natl Acad Sci USA,* 94**,** 8380-8386.

DURET, L. & MOUCHIROUD, D. (2000) Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. *Mol Biol Evol,* 17**,** 68-74.

EBERSBERGER, I., METZLER, D., SCHWARZ, C. & PÄÄBO, S. (2002) Genome wide comparison of DNA sequences between humans and chimpanzees. *Am J Hum Genet,* 70**,** 1490-1497.

ELLEGREN, H. (2007) Characteristics, causes and evolutionary consequences of male-biased mutation. *Proc Biol Sci,* 274**,** 1-10.

FARKASH-AMAR, S., LIPSON, D., POLTEN, A., GOREN, A., HELMSTETTER, C., YAKHINI, Z. & SIMON, I. (2008) Global organization of replication time zones of the mouse genome. *Genome Research,* 18**,** 1562-1570.

GAFFNEY, D. J. & KEIGHTLEY, P. D. (2005) The scale of mutational variation in the murid genome. *Genome Research,* 15**,** 1086-1094.

GOETTING-MINESKY, M. P. & MAKOVA, K. D. (2006) Mammalian male mutation bias: impacts of generation time and regional variation in substitution rates. *Journal of Molecular Evolution,* 63**,** 537-544.

GREEN, P., EWING, B., MILLER, W., THOMAS, P. J., PROGRAM, N. C. S. & GREEN, E. D. (2003) Transcription-associated mutational asymmetry in mammalian evolution. *Nat Genet,* 33**,** 514-517.

HIRATANI, I., RYBA, T., ITOH, M., YOKOCHI, T., SCHWAIGER, M., CHANG, C.-W., LYOU, Y., TOWNES, T. M., SCHÜBELER, D. & GILBERT, D. M. (2008) Global reorganization of replication domains during embryonic stem cell differentiation. *PLoS Biol,* 6**,** e245.

HOLMQUIST, G. P. (1987) Role of replication time in the control of tissue-specific gene expression. *Am J Hum Genet,* 40**,** 151-173.

HURST, L. D. & ELLEGREN, H. (1998) Sex biases in the mutation rate. *Trends Genet,* 14**,** 446-452.

HURST, L. D. & WILLIAMS, E. J. (2000) Covariation of GC content and the silent site substitution rate in rodents: implications for methodology and for the evolution of isochores. *Gene,* 261**,** 107-114.

JACKSON, D. A. & POMBO, A. (1998) Replicon clusters are stable units of chromosome structure: evidence that nuclear organization contributes to the efficient activation and propagation of S phase in human cells. *The Journal of Cell Biology,* 140**,** 1285-1295.

KARNANI, N., TAYLOR, C., MALHOTRA, A. & DUTTA, A. (2007) Pan-S replication patterns and chromosomal domains defined by genome-tiling arrays of ENCODE genomic areas. *Genome Research,* 17**,** 865-876.

KOÇ, A., WHEELER, L. J., MATHEWS, C. K. & MERRILL, G. F. (2004) Hydroxyurea arrests DNA replication by a mechanism that preserves basal dNTP pools. *J Biol Chem,* 279**,** 223-230.

LERCHER, M. J., CHAMARY, J.-V. & HURST, L. D. (2004) Genomic regionality in rates of evolution is not explained by clustering of genes of comparable expression profile. *Genome Research,* 14**,** 1002-1013.

LERCHER, M. J., WILLIAMS, E. J. & HURST, L. D. (2001) Local similarity in evolutionary rates extends over whole chromosomes in human-rodent and mouse-rat comparisons: implications for understanding the mechanistic basis of the male mutation bias. *Mol Biol Evol,* 18**,** 2032-2039.

LI, W. H., YI, S. & MAKOVA, K. (2002) Male-driven evolution. *Curr Opin Genet Dev,* 12**,** 650-656.

MAJEWSKI, J. (2003) Dependence of mutational asymmetry on gene-expression levels in the human genome. *Am J Hum Genet,* 73**,** 688-692.

MAKOVA, K. D. & LI, W.-H. (2002) Strong male-driven evolution of DNA sequences in humans and apes. *Nature,* 416**,** 624-626.

MALCOM, C. M., WYCKOFF, G. J. & LAHN, B. T. (2003) Genic mutation rates in mammals: local similarity, chromosomal heterogeneity, and X-versus-autosome disparity. *Mol Biol Evol,* 20**,** 1633-1641.

MALÍNSKY, J., KOBERNA, K., STANĚK, D., MASATA, M., VOTRUBA, I. & RASKA, I. (2001) The supply of exogenous deoxyribonucleotides accelerates the speed of the replication fork in early S-phase. *J Cell Sci,* 114**,** 747-750.

MARTOMO, S. A. & MATHEWS, C. K. (2002) Effects of biological DNA precursor pool asymmetry upon accuracy of DNA replication in vitro. *Mutat Res,* 499**,** 197-211.

MATASSI, G., SHARP, P. M. & GAUTIER, C. (1999) Chromosomal location effects on gene sequence evolution in mammals. *Curr Biol,* 9**,** 786-791.

MATHEWS, C. K. (2006) DNA precursor metabolism and genomic stability. *FASEB J,* 20**,** 1300-1314.

MCVEAN, G. T. & HURST, L. D. (1997) Evidence for a selectively favourable reduction in the mutation rate of the X chromosome. *Nature,* 386**,** 388-392.

MIYATA, T., HAYASHIDA, H., KUMA, K. & YASUNAGA, T. (1987) Male-driven molecular evolution demonstrated by different rates of silent substitutions between autosome-and sex chromosome-linked genes. *Proceedings of the Japan Academy. Ser. B: Physical and Biological Sciences,* 63**,** 327-331.

MUGAL, C. F., VON GRÜNBERG, H.-H. & PEIFER, M. (2009) Transcription-induced mutational strand bias and its effect on substitution rates in human genes. *Mol Biol Evol,* 26**,** 131-142.

PINK, C. J., SWAMINATHAN, S. K., DUNHAM, I., ROGERS, J., WARD, A. & HURST, L. D. (2009) Evidence that replication-associated mutation alone does not explain between-chromosome differences in substitution rates. *Genome Biology and Evolution,* 2009**,** 13-22.

POLAK, P. & ARNDT, P. F. (2008) Transcription induces strand-specific mutations at the 5' end of human genes. *Genome Research,* 18**,** 1216-1223.

SANDSTEDT, S. A. & TUCKER, P. K. (2005) Male-driven evolution in closely related species of the mouse genus *Mus*. *Journal of Molecular Evolution,* 61**,** 138-144.

SHIMMIN, L. C., CHANG, B. H. & LI, W. H. (1993) Male-driven evolution of DNA sequences. *Nature,* 362**,** 745-747.

SMITH, N. G. & HURST, L. D. (1999) The causes of synonymous rate variation in the rodent genome. Can substitution rates be used to estimate the sex bias in mutation rate? *Genetics,* 152**,** 661-673.

STAMATOYANNOPOULOS, J. A., ADZHUBEI, I., THURMAN, R. E., KRYUKOV, G. V., MIRKIN, S. M. & SUNYAEV, S. R. (2009) Human mutation rate associated with DNA replication timing. *Nat Genet,* 41**,** 393-395.

TAMURA, K. & KUMAR, S. (2002) Evolutionary distance estimation under heterogeneous substitution pattern among lineages. *Mol Biol Evol,* 19**,** 1727-1736.

TOUCHON, M., ARNEODO, A., D'AUBENTON-CARAFA, Y. & THERMES, C. (2004) Transcription-coupled and splicing-coupled strand asymmetries in eukaryotic genomes. *Nucleic Acids Res,* 32**,** 4969-4978.

TOUCHON, M., NICOLAY, S., ARNEODO, A., D'AUBENTON-CARAFA, Y. & THERMES, C. (2003) Transcription-coupled TA and GC strand asymmetries in the human genome. *FEBS Lett,* 555**,** 579-582.

WEBSTER, M. T., SMITH, N. G. C., LERCHER, M. J. & ELLEGREN, H. (2004) Gene expression, synteny, and local similarity in human noncoding mutation rates. *Mol Biol Evol,* 21**,** 1820-1830.

WEDDINGTON, N., STUY, A., HIRATANI, I., RYBA, T., YOKOCHI, T. & GILBERT, D. M. (2008) ReplicationDomain: a visualization tool and comparative database for genome-wide replication timing data. *BMC Bioinformatics,* 9**,** 530.

WOLFE, K. H., SHARP, P. M. & LI, W. H. (1989) Mutation rates differ among regions of the mammalian genome. *Nature,* 337**,** 283-285.

WOODFINE, K., FIEGLER, H., BEARE, D. M., COLLINS, J. E., MCCANN, O. T., YOUNG, B. D., DEBERNARDI, S., MOTT, R., DUNHAM, I. & CARTER, N. P. (2004) Replication timing of the human genome. *Hum Mol Genet,* 13**,** 191-202.

YAFFE, E., FARKASH-AMAR, S., POLTEN, A., YAKHINI, Z., TANAY, A. & SIMON, I. (2010) Comparative analysis of DNA replication timing reveals conserved large-scale chromosomal architecture. *PLoS Genet,* 6**,** e1001011.

Chapter 4. **A Gender-Specific Relationship Between Replication Time and Recombination Rate: Implications for Understanding the Determinants of K_i and GC Content**

Catherine J. Pink and Laurence D. Hurst

Based on a paper and supplementary information published at:

PLoS One (2011). 6(9): e24480

4.1 Introduction

Numerous studies, including those in Chapter 2, have now shown that neutral sites in mammals evolve faster in domains of high recombination (Pink *et al*., 2009, Lercher and Hurst, 2002, Tyekucheva *et al*., 2008, Hellmann *et al*., 2003, Perry and Ashworth, 1999). That the strength of the correlations reported tends to be weak likely reflects inexact measures of recombination rate, which have been shown to be fast evolving (Ptak *et al*., 2005, Dumont *et al*., 2011). Two possible explanations for this relationship have been proposed: First, it has been suggested that the recombination process is mutagenic (Magni and Von Borstel, 1962, Magni, 1963, Strathern *et al*., 1995). Alternatively, even if recombination is not mutagenic, biased gene conversion can promote the fixation of neutral mutations and can increase rates of evolution that are not at equilibrium (Piganeau *et al*., 2002). Due to biases in the mismatch repair process (Marais, 2003), the latter process tends to favour fixation of G/C over A/T and thus has also been suggested as a mechanism for the origin or maintenance of isochores (Meunier and Duret, 2004, Duret and Galtier, 2009 and references therein).

Comparably, as shown in Chapter 3, rates of evolution have been shown to increase across S-phase (Pink and Hurst, 2010, Stamatoyannopoulos *et al*., 2009, Chen *et al*., 2010), while GC rich sequences tend to be early replicating (Pink and Hurst, 2010,

Woodfine *et al*., 2004, Costantini and Bernardi, 2008). Given this, GC rich sequences might therefore be expected to evolve slowly. Indeed, the analyses in Chapters 2 and 3 showed this to be the case with negative relationships between GC content and intronic rates of evolution. Chapter 3 also demonstrated that the relationship between replication time and intronic rates of evolution was not owing to GC content, but that the lower intronic substitution rates of GC rich sequences could be explained by GC rich sequences being early replicating, and thus slow evolving.

What has not yet been established is the extent to which these two variables, replication timing and recombination rate, are independent predictors of neutral rates of evolution. Given that *a priori* GC rich sequences tend to be a) early replicating and b) highly recombining, it might be expected that recombination rates are highest when replication is earliest in the cell cycle. Given that early replication is associated with low rates of evolution, whereas recombination is associated with higher substitution rates, the two factors might therefore cancel each other out. This raises the possibility that the magnitude of the impact of both replication time and of recombination rate on rates of evolution might have been underestimated when either factor is considered in isolation, even if the underlying effects are relatively strong, owing to masking effects. If so, this would necessitate a need to control one for the other. As this question was not addressed in Chapter 3, it is instead examined here, both at the genic level and also with regard to the enigmatic between-autosome variation in neutral rates (Pink *et al*., 2009, Lercher *et al*., 2001, Malcom *et al*., 2003).

An increasing body of evidence suggests that the effect of recombination on weak-to-strong (A/T to G/C) substitutions correlates more strongly with rates in males than in females (Dreszer *et al*., 2007, Duret and Arndt, 2008, Tyekucheva *et al*., 2008, Webster *et al*., 2005, Berglund *et al*., 2009, Galtier *et al*., 2009). The reasons why this might be have not yet been elucidated, although a mechanistic difference in meiotic recombination has been suggested (Galtier *et al*., 2009). Given the potential importance of sex-specific recombination rates, this study considered not just sex-

averaged recombination rates but repeated all analyses using both male- and female-specific recombination rates.

With the inclusion of sex-specific recombination rates, this analysis differed from that of Chen *et al*. (2010) who argued that the effect of replication timing on neutral evolutionary rates is not explained by recombination. This group, however, only examined sex-averaged rates. This analysis also differed from that of Clément and Arndt (2011) who noticed that GC content in rodents is well predicted by male-specific recombination rates but not by female-specific ones and therefore chose to ignore further consideration of female recombination as a potentially important cause of GC content.

The use of rodent replication time and recombination rate data is not straightforward. These data are typically not supplied as genic tracks from major repositories, but as supplementary or standalone files containing raw data values pertaining to SNP or array probe positions. As such, these data sets are often not updated to current assemblies of the genome and so assignment of position-based data to genes requires careful consideration. The analyses presented here utilise data sets curated under what might be considered improved methodologies, representing improvements over those used in previous chapters. As such, the scope of this chapter was expanded to test whether both prior and novel conclusions are robust to changes in methodology.

4.2 Methods

4.2.1 Estimating intronic substitution rates

Using the same methodologies as applied in Chapters 2 and 3, a new autosomal intronic substitution rate data set was generated for these analyses. These methods are described in detail in sections 2.2.1 to 2.2.6, including the filter for clusters of conserved bases, potentially indicative of hidden functional sites, described in section 2.2.4. Introns of the same gene were concatenated, the rate of intronic divergence (K_i) estimated and corrected for multiple hits according to the model of Tamura and Kumar (2002). The autosomal distributions of these new data are shown in Figure 4.1.

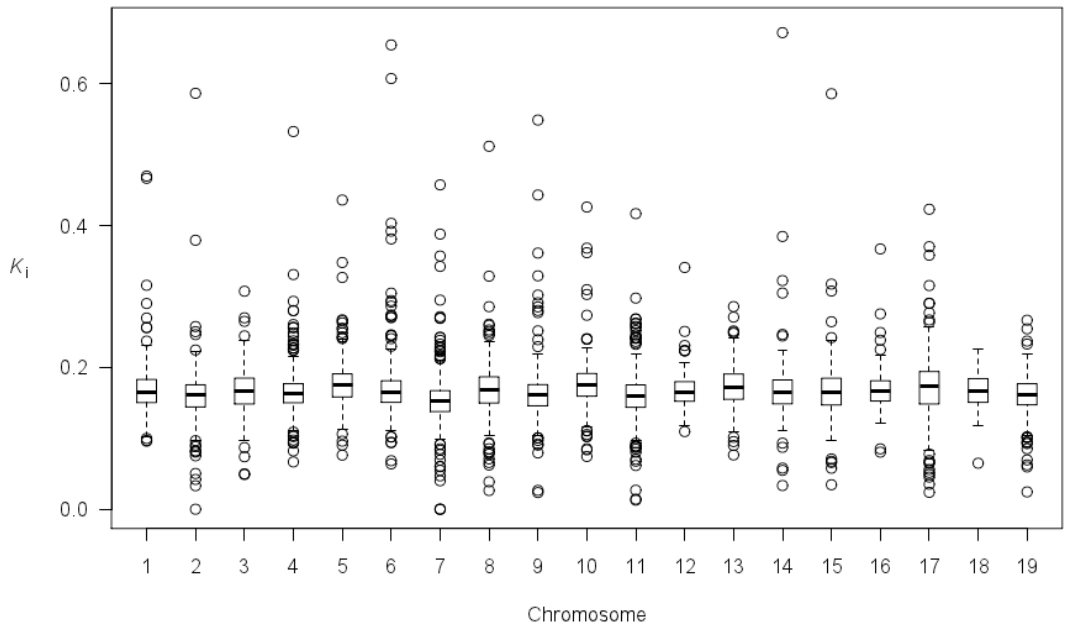


Figure 4.1: Boxplot showing the distribution of genic substitution rates, measured at intronic sites (K_i), across each autosome. Note that one outlier for chromosome 10 ($K_i = 0.918$) is not shown.

4.2.2 Estimating GC content

Mouse GC content was calculated directly from genomic sequences at intronic sites using both unmasked sequences and repeat masked sequences – the latter to control for the possible influence of AT rich transposable element insertions. Genomic sequence files for the mouse genome mm9 (NCBI build 37, July 2007) were obtained from the UCSC table browser located at http://genome.ucsc.edu/ (Karolchik *et al*., 2004). A second duplicate set of gene sequences were downloaded with repeat sequences masked to N. Dubious RefSeqs that either were present in more than one copy, were found to be located on random or multiple chromosomes, that were not located on a single strand, or that were intron-less were identified and removed from the analysis. Coding sequences that either did not begin and end with correct start and termination codons, that consisted of incomplete codons or that contained premature stop codons were identified and all intronic sequences for these RefSeqs were purged from the analysis. For both repeat masked and unmasked intronic sequences, 30 bp were removed from both ends of each intron to control for the possible influence of conserved splice sites (Touchon *et al*., 2004). First introns were also removed, these known to be unusually slow evolving (Keightley and Gaffney,

2003, Chamary and Hurst, 2004). Remaining intronic sequences were then concatenated by RefSeq. Counts of each base (A, T, C, G and N) were then made from which GC content (GC) was calculated as [(G + C) / (A + T + G + C)].

4.2.3 The rearrangement index

The methodology used to calculate each autosome's rearrangement index (RI) using the newly curated substitution rate data set was the same as that used in Chapters 2 and 3, described in detail in section 2.2.11. However, in contrast to these analyses the number of bootstrap randomisations used for each autosome was increased to 10,000.

4.2.4 Assaying replication time

Replication times in *Mus musculus* were determined by Hiratani *et al*. (2008). As described in Chapter 3, four replication timing data sets were available. Three were derived from separate embryonic stem cell lines (ECSs). Inclusion of a fourth data set derived from induced pluripotent stem cells (iPS) was justified in Chapter 3 (Pink and Hurst, 2010) and so was again included. These data sets were downloaded in files RD_TT2ESCave_Sm300_081128.txt, RD_iPSave_Sm300_081128.txt, RD_D3ESCave_Sm300_081128.txt and RD_46CESCave_Sm300_081128.txt from the ReplicationDomain website (Weddington *et al*., 2008). Array probe positions were converted from mouse build mm8 (NCBI build 36) to build mm9 (NCBI build 37) using the stand alone UCSC liftOver tool and associated chain file mm8ToMm9.over.chain. This method was considered preferable to that used for the analyses in Chapter 3 as it allowed for possible cases whereby an individual marker's position had been relocated from within a gene to inter-genic sequence or vice-versa. All probes located within the limits of the coding sequence of a RefSeq were then identified. Of the 21,471 RefSeqs, 14,881 were assigned sufficient replication times to be able to test for normality of distribution for the RefSeq. Kolmogorov-Smirnov tests showed that replication times of 5,126 RefSeqs (35.5% of those tested) were normally distributed while 9,755 (65.6% of those tested) had skewed distributions. Median replication times were therefore assigned to each RefSeq. It should be noted that use of mean replication times did not qualitatively alter the findings (see Supplementary Tables 4.1, 4.2 and 4.3).

4.2.5 Methods to estimate the local recombination rate

In contrast to the analyses in Chapter 2 that utilised recombination rates in rat, here recombination rates in mouse were used. This enabled comparison of the relative contributions of recombination and replication time to rates of evolution in a single species. The genetic map used was originally determined by Shifman *et al*. (2006), derived from a large heterogeneous mouse population descended from eight inbred strains. Cox *et al*. (2009), having identified two methodological problems with the Shifman genetic map, subsequently updated this data set and incorporated SSLP markers from other genetic maps to generate a revised standard genetic map for the mouse. The map consists of 10,195 SNPS at an average density of 258 Kb (99% of SNP intervals <500 Kb, 81.2% <250 Kb) and is based on 3,546 meioses. This revised genetic map was therefore used for this analysis. The genetic map was downloaded from http://cgd.jax.org/mousemapconverter/Revised_HSmap_SNPs.csv - Mouse Map Data (Base Pair to centimorgan mapping). SNP positions had already been updated to the current mouse build mm9 (NCBI build 37). In addition to the SNP ID, the chromosome and physical base pair position of the SNP, this file contained three genetic maps: a male-specific map, a female-specific map and a sex-averaged map. Assignment of recombination rates to RefSeqs was performed using a number of alternative methodologies:

4.2.5.1 Chromosomal recombination rates are generally calculated from the most proximal and distant markers. Doing so captures all recombination events along the chromosome. Application of a similar methodology to individual RefSeqs involved identification of the two flanking SNPs. The physical and genetic positions of these markers could then be used to calculate the recombination rate of the intervening region in which the RefSeq was located (Figure 4.2). The median distance between the edge of a gene and the flanking marker was 155346.5 bp, indicating that this methodology estimated genic recombination rates over an approximate 300 Kb window.

4.2.5.2 Human recombination rates, such as the deCODE, Marshfield and Genethon genetic maps, are available as additional tracks on the UCSC genome

browser. These are essentially weighted averages, whereby the recombination rate between immediately flanking markers is calculated and, assuming a linear genetic distance between markers, each base within the interval is assigned the recombination rate. 1Mb windows are then assigned recombination rates based on the average rate of the bases contained within the window. A similar method was therefore applied to genes, albeit without smoothing over 1Mb windows. RefSeqs were assigned mean recombination rates weighted by the base pair overlap of the marker interval with the gene (Figure 4.2). This was, in effect, the same as assigning each base pair within the gene a recombination rate and then taking a mean across all base pairs. A 'weighted median' was also calculated by assigning each base pair within the gene a recombination rate and then taking a median across all base pairs, since the per-base pair recombination rates of over 1000 genes had skewed distributions.

4.2.5.3 A method similar to that applied to the assignment of replication times to each RefSeq was also used. Here, for each chromosome the recombination rate between every neighbouring pair of SNPs was calculated. Each SNP interval that overlapped with a given RefSeq was identified and the average mean and median recombination rate of these intervals was taken (Figure 4.2). Note that for genes that lacked internal SNPs, this resulted in the same genic recombination rate as for method 4.2.5.1.

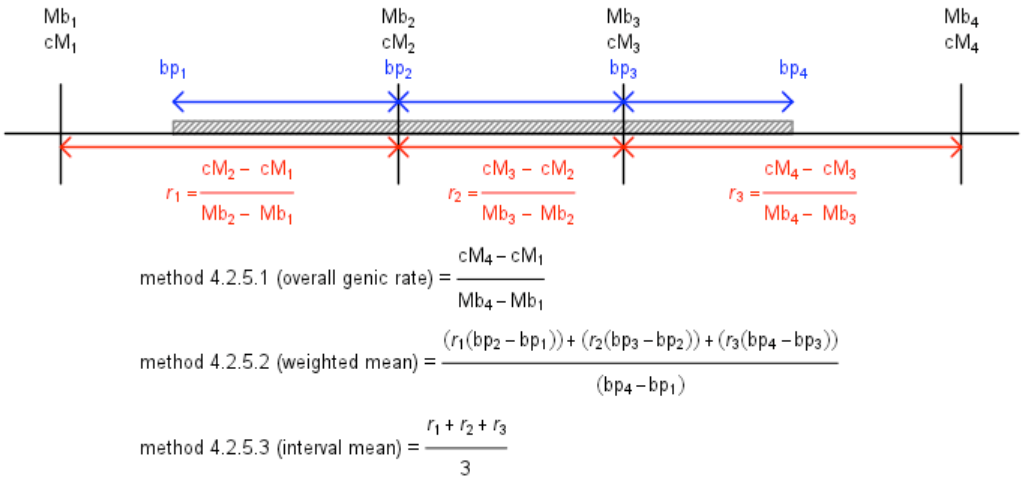


Figure 4.2: Representation of the methods used to calculate gene-focused recombination rates (methods 4.2.5.1, 4.2.5.2 and 4.2.5.3). Note that this diagram is for descriptive purposes only and is not to scale. For simplicity, only calculations for mean rates are shown. The grey region is a gene. Vertical black lines represent four SNP markers with physical (Mb) and genetic (cM) positions. Blue arrows represent the base pairs of the gene overlapping with each intervening interval. In red are recombination rates (r_x) between pairs of neighbouring markers.

4.2.5.4 To reduce noise, smoothing techniques were also applied. Two methods of smoothing were used and in each case, both means and medians were used, thus giving four smoothed rates. Firstly, all markers within a 2Mb window of the flanking interval were identified (1Mb in each direction from the 5' SNP). Recombination rates between each pair of markers were calculated, again assuming a linear genetic distance between markers. The average recombination rate of all these marker intervals was taken and assigned to the focal interval (denoted average-smoothed[1] in the text). Secondly, in addition to the focal interval, these 2Mb averaged recombination rates were assigned to every interval within the 2Mb window. Once this process had been repeated using all intervals as a focal point for the 2Mb smoothing, the average of all smoothed rates assigned to a window was taken (denoted average-smoothed[2] in the text). Finally, these four smoothed rates were assigned to genes using the same technique as described in method 4.2.5.3 (Figure 4.3).

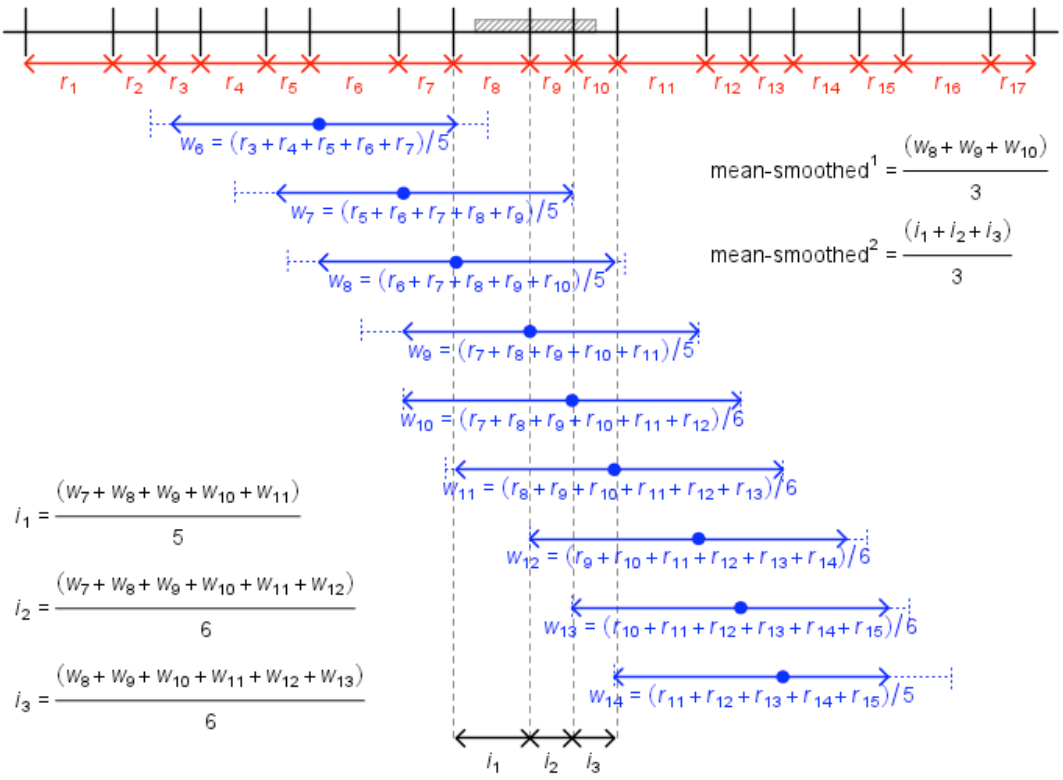


Figure 4.3: Representation of methods used to calculate smoothed recombination rates (method 4.2.5.4). Note that this diagram is for descriptive purposes only and is not to scale. For simplicity, only calculations for mean rates are shown. The grey region is a gene. Vertical black lines are SNP markers. In red are recombination rates between pairs of neighbouring markers (r_x). Dashed blue lines represent 1Mb windows either side of a focal SNP. Solid blue arrows represent all intervals within this window, over which recombination rates are averaged (w_x, smoothing method 1). For three intervals, averages of all window averages covering the interval are shown (i_x, smoothing method 2).

These alternative methodologies were applied to both the sex-averaged, male-specific and female-specific data. Every statistical analysis that included recombination rate as a parameter was repeated using every method described.

4.2.6 Data set dimensions

For the analyses presented here, the final data set was purged of all sex-linked RefSeqs. In addition, only RefSeqs that had been assigned data for all variables of interest - intronic substitution rates (K_i), GC content (GC), timing of replication (RT), and recombination rate (RR) - were retained, thus ensuring that the sample

size, and therefore statistical power, was comparable across all analyses. The resulting data set comprised 3,549 genes.

For all genic data sets, Kolmogorov-Smirnov tests were applied, showing that data were skewed and could not be normalised. Similarly, Kolmogorov-Smirnov tests performed on data assigned to individual autosomes showed that all data types were also skewed. As such, for analyses of between-autosomal variation, the median autosomal value for each data type was taken. To these autosomal medians, the overall recombination rate between the most proximal and distal markers on the chromosome, plus the rearrangement indices were added. Finally, for each data type the distributions of the 19 autosomal values were found to be normally distributed, thus enabling the use of parametric tests for analyses at the autosomal level.

4.2.7 Calculation of partial spearman correlations

Partial Spearman's correlations between *x* and *y*, controlling for *z* ($\rho_{xy.z}$), were calculated as previously described in Section 3.2 7.

4.3 Results

4.3.1 Repeat masked GC_i used as an approximation of stationary GC

Recombination has been shown to correlate more strongly with the stationary GC content (GC*) than with current GC content (Duret and Arndt, 2008, Duret and Galtier, 2009). Calculation of GC* is dependent on polarisation of substitutions based on the ancestral state, achieved by reference to an outgroup. At the conception of this work, no rodent assembly was considered to have high enough coverage to enable this to be done for mouse-rat substitutions and complete non-rodent assemblies such as human were considered too distant a relative to enable accurate assignment of the ancestral state. However, Clément and Arndt (2011) recently suggested that the use of human as an outgroup for mouse-rat substitutions generated similar results to those obtained using the current drafts of guinea pig (*Cavia porcellus*) or kangaroo rat (*Dipodomys ordii*). In light of these findings, calculation of GC* should be possible for future analyses.

That GC* typically differs from current GC content suggests that mammalian genomes are not yet at mutational equilibrium. One reason why this might be is the insertion of young transposable elements (TEs) that have atypical GC content. In an attempt to remove such sites and therefore provide a closer approximation to GC*, GC content was computed using repeat masked intronic sequences, whereby repeats with <40% divergence were masked to 'N'. Repeat masked and unmasked GC_i were found to strongly covary (Spearman's $\rho = 0.983$, $P < 2.2 \times 10^{-16}$; $n = 18775$;).

Based on the nucleotide composition and differential location of different classes of TEs (specifically L1 and Alu insertions), Duret and Hurst (2001) showed that in humans, insertion of TEs in AT rich introns would elevate GC content whereas in GC rich isochores, insertion of TEs would reduce the intronic GC content. Consistent with Duret and Hurst's (2001) prediction, Figure 4.4 shows that in GC poor introns, unmasked sequences did indeed have a higher GC content than repeat masked sequences. Similarly, in GC rich introns, the repeat masked GC_i was higher than that of the unmasked introns. This suggested that repeat masking removed some sites that were unlikely to be at mutational equilibrium.

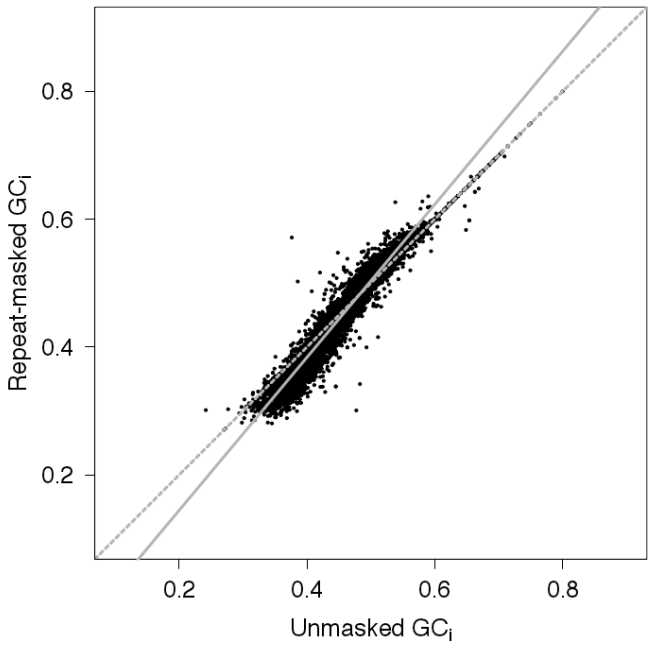


Figure 4.4: Covariance of unmasked and repeat-masked intronic GC content. The dashed line represents $x = y$. The solid line is the orthogonal regression where Repeat-masked GC_i = -0.095 + 1.196949Unmasked GC_i.

4.3.2 A sex-specific relationship between replication time and recombination rate at the genic level

Initially, two sets of questions were asked: First, was it robustly found that replication time and the local recombination rate both correlate with the intronic substitution rate? Second, was it true that recombination and replication time covaried as expected? If the second were true then the former results would need to be analysed under a covariate controlled model.

With regard to the first issue, the previously observed relationship between replication timing and rates of intronic evolution, shown in Chapter 3, was confirmed in the new data set (Spearman's $\rho = -0.081$, $P = 1.35 \times 10^{-6}$). Note again that because of how the replication timing data was structured, an increase in any parameter as S-phase proceeds yields a negative correlation and *vice versa*. The relationship between recombination rates and intronic substitution rates was more complex, being sensitive to both gender and methodology. In general, all recombination rate data sets that involved an element of smoothing resulted in stronger correlations with K_i than the gene-focused data sets such as overall rates, weighted, base pair and interval averages (Table 4.1, pages 131-132). For smoothed rates, the magnitude of the relationship was similar to that observed for replication times (for mean-smoothed[2] sex-averaged recombination rates Spearman's $\rho = 0.1$, $P = 2.39 \times 10^{-9}$) whereas for unsmoothed rates, the strength of the relationship was approximately half that for replication times (for overall sex-averaged recombination rates Spearman's $\rho = 0.045$, $P = 0.0073$).

Interestingly, the relationship between substitution rates and recombination appeared to be driven by recombination in females: all female-specific recombination rates showing significant positive correlations with K_i, whereas for male-specific recombination rates, correlation coefficients for smoothed data sets were approximately half the magnitude of those for females and for gene-focused data sets no significant relationships were observed (Table 4.1). This was surprising, as weak-to-strong substitutions associated with GC biased gene conversion (gBGC) in primates have been found to covary more strongly with male-specific recombination

rates (Webster *et al*., 2005, Dreszer *et al*., 2007, Tyekucheva *et al*., 2008, Berglund *et al*., 2009, Duret and Arndt, 2008, Galtier *et al*., 2009).

As to the second question of whether timing of replication and recombination rates covaried, unexpectedly no consistent relationship was observed for sex-averaged recombination rates, with both increasing and declining rates associated with sequences that replicate later during S-phase (Table 4.1). Closer examination suggested that this result reflected differences between males and females (Figure 4.5). Female recombination rates were consistently found to be higher in regions that replicate later during S-phase, irrespective of smoothing (for overall female recombination rates Spearman's $\rho = -0.076$, $P = 6.34 \times 10^{-6}$, Table 4.1). In contrast, genes that replicated later were found to have significantly lower male-specific recombination rates for some methodologies (e.g. for mean-smoothed[2] male recombination rates Spearman's $\rho = 0.138$, $P = 1.21 \times 10^{-16}$, Table 4.1) whereas for other measures no relationship was observed (e.g. for overall male recombination rates Spearman's $\rho = 0.025$, $P = 0.135$, Table 4.1).

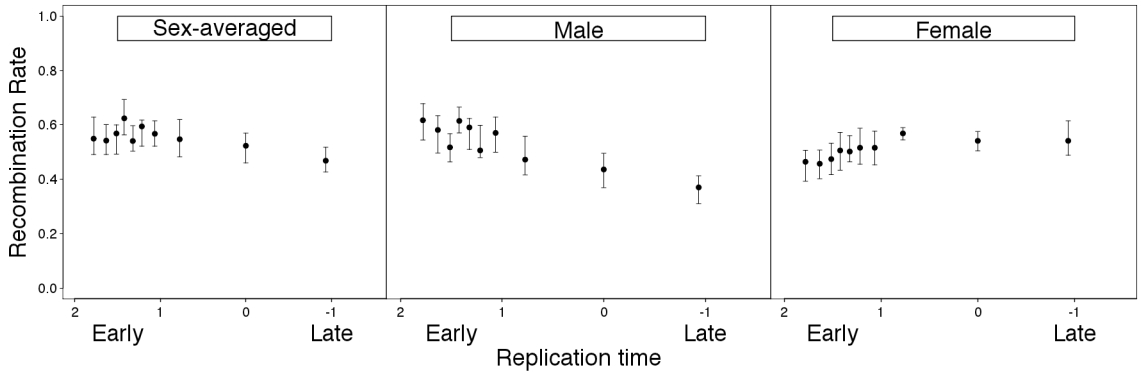


Figure 4.5: Relationship between replication time and sex-averaged, male-specific and female-specific recombination rates. Data shown are mean-smoothed[2] data binned by median replication time where points are the median of each bin ± 95% confidence intervals.

4.3.3 Weak interference between replication timing and sex-specific recombination rates in determining intronic substitution rates

Given this result, it was therefore necessary to ask whether the high substitution rate of late replicating sequence was due to it having high recombination rates in females and *vice versa*. Similarly, it was necessary to ask wither the impact of male

recombination on rates of evolution have been underestimated as male-specific recombination rates were low where the effect of replication was strongest.

Indeed, controlling for female recombination was found to reduce the strength of the relationship between K_i and replication time. This was the case for all female-specific data sets (for the uncontrolled analysis Spearman's $\rho = -0.081$, $P = 1.35 \times 10^{-6}$; controlling for overall female recombination partial Spearman's $\rho = -0.078$, $P = 0.001$, Table 4.1), although the effect appeared quite modest. Similarly, controlling for replication time reduced the strength of the relationship between intronic substitution rate and all measures of female-specific recombination rate (for the uncontrolled relationship between K_i and overall female recombination, Spearman's $\rho = 0.044$, $P = 0.0090$; controlling for replication time partial Spearman's $\rho = 0.038$, $P = 0.013$, Table 4.1).

In contrast, the higher male-specific recombination rates of early replicating sequences might have masked the impact of replication time on rates of evolution and *vice versa*. When controlling for male recombination the magnitude of the relationship between K_i and replication time might therefore have been expected to increase. Controlling for gene-focused measures of male recombination did not affect the covariance between replication time and K_i (for the uncontrolled analysis Spearman's $\rho = -0.081$, $P = 1.35 \times 10^{-6}$; controlling for overall male recombination partial Spearman's $\rho = -0.081$, $P = 0.001$, Table 4.1). However, a slight increase in the strength of this relationship was indeed observed when controlling for smoothed measures of male recombination and was greatest for those that had shown the strongest positive covariance between recombination rate and replication time (controlling for mean-smoothed^2 male recombination rates, partial Spearman's $\rho = -0.09$, $p = 0.001$, Table 4.1). Likewise, the lack of any relationship between K_i and all gene-focused measures of male-specific recombination was not affected by controls for replication time (P remained >0.05 for all, Table 4.1). However, a slight increase in the strength of the relationship between K_i and all smoothed measures of male recombination was observed (for the uncontrolled relationship between K_i and mean-smoothed^2 male recombination, Spearman's $\rho = 0.058$, $P = 0.0005$; controlling for replication time, partial Spearman's $\rho = 0.07$, $P = 0.001$, Table 4.1).

Together, these results suggested that in estimating the impact of either timing of replication or recombination on the rate of neutral substitutions it is thus helpful, at the genic level, to perform a covariate controlled analysis. However, as the correction is small, this isn't essential.

4.3.4 Autosomal rates of evolution were better predicted by replication time than by recombination rate

The above analyses considered relationships at the genic level, but what about variation at the autosomal level? As discussed in Chapter 2, for as yet unidentified reasons, more highly rearranged mouse autosomes have been found to have higher substitution rates (for the new data set Pearson's $r = 0.761$, $P = 0.0002$; least squares linear regression $r^2 = 0.579$, $P = 0.0002$). As such, the extent of inter-autosomal rearrangement should be considered alongside any other parameters under investigation as predictors of between-autosomal variation in K_i. To account for this a residuals test was therefore used whereby the residuals from the above regression were predicted by variation in the parameter of interest.

In Chapter 3 it was shown that although replication time alone was unable to explain between-autosomal variation in rates of evolution, it was a significant predictor of this residual variation. These findings were confirmed in the new data set: Although autosomal substitution rates did not covary with autosomal replication times (Pearson's $r = -0.272$, $P = 0.26$), residual variation in median K_i not explained by the rearrangement index could be predicted by differences in median timing of replication ($r^2 = 0.237$, $P = 0.034$), whereby earlier replicating autosomes had lower substitution rates than predicted by the rearrangement index and later replicating autosomes evolved faster than would be predicted by extent of rearrangement. When combined in a multiple least squares linear regression, rearrangement index and replication time could together explain around 68% of inter-autosomal variation in K_i ($r^2 = 0.679$, $P = 0.0001$) and both parameters were significant predictors in this model ($P = 4.89 \times 10^{-5}$ for rearrangement index; $P = 0.04$ for replication time).

When autosomal recombination rates were subjected to a similar analysis, they too were found not to covary with autosomal rates of intronic evolution (for overall sex-averaged recombination rates Pearson's r = -0.182, P = 0.457, Supplementary Table 4.2). However, application of the same residuals test showed that unlike replication time, residual variation from the regression of K_i against rearrangement index could not be accounted for by autosomal recombination rates (for overall sex-averaged recombination rates r^2 = 0.018, P = 0.581, Supplementary Table 4.3). Further, the predictive power of the model to explain autosomal rates of evolution by the rearrangement index was only marginally increased by the inclusion of recombination rates (r^2 = 0.584, P = 0.00090) and recombination rate was not a significant predictor in the model (P = 0.00047 for rearrangement index; P = 0.673 for recombination rate). These findings were all robust to the use of alternative methods of assigning autosomal recombination rates and to the use of either male- or female-specific recombination rates (Supplementary Table 4.3).

That replication timing was a somewhat stronger covariate of K_i than recombination rate, particularly at the autosomal level, might in part have been explained by the impact of extensive genomic rearrangements in the mouse lineage (Ramsdell *et al.*, 2008). As homologous regions have highly conserved replication times, this suggests that as sequences move around the genome, they tend to take their replication times with them (Chen *et al.*, 2010, Farkash-Amar *et al.*, 2008, Yaffe *et al.*, 2010). In contrast, the relocation of rodent centromeres from a metacentric to a telocentric location has reduced the number of chromosome arms and, based on the requirement for at least one chiasma per arm, reduced the overall recombination rate of each autosome (Jensen-Seaman *et al.*, 2004). Further, recombination hotspots are known to be short lived (Ptak *et al.*, 2005, Dumont *et al.*, 2011). As such, while substitution rates and GC content are the product of processes occurring over long periods of time, the current replication time of a given sequences is more likely to reflect that to which it has been exposed to ancestrally than is the case for current recombination rates.

4.3.5 GC content is better predicted by replication timing than by recombination rates.

Current thinking suggests that the isochore structure of mammalian genomes is the result of recombination-associated biased gene conversion and that this process has a more profound effect in the male than in the female germline. However, early replicating sequences are known to be GC rich. Indeed, more generally, a relationship between isochore boundaries and replication time boundaries is well described both on local and genomic scales (Woodfine *et al*., 2004, Costantini and Bernardi, 2008, Watanabe *et al*., 2002, Schmegner *et al*., 2005, Schmegner *et al*., 2007). Was then the local GC content better predicted by replication timing than by recombination rate and did this help to explain why male recombination, rather than female recombination appeared to be relevant?

It was striking that although the strength of the relationship between genic GC content and replication time was lower than previously observed (Spearman's ρ = 0.293, $P = 5.34 \times 10^{-71}$ versus Spearman's ρ = 0.315, Chapter 3), early replicating sequences being more GC rich, timing of replication was a stronger correlate of GC content than were all measures of recombination rate (Spearman's ρ = 0.067, $P = 6.44 \times 10^{-5}$ for overall sex-averaged recombination, Table 4.1). Although the direction of the genic relationship was robust with highly recombining genes consistently having higher GC contents, the strength of the relationship was sensitive to gender: male-specific recombination rates being a stronger covariate of GC content than female-specific rates (Table 4.1). Methodology was also an important factor in determining the nature of the relationship. Gene-focused data sets were generally qualitatively similar. In contrast, the method of smoothing generated contrasting results: Use of medians to smooth both male and female recombination rates negated the significance of the relationship whilst for both genders the strongest correlate of GC content was mean-smoothed[2] recombination rates (Table 4.1).

At the autosomal level, the contrast between replication timing and recombination rate as predictors was even more pronounced, with higher autosomal GC content correlating strongly with earlier autosomal replication (Pearson's r = 0.832, P = 9.83

x 10^{-6}) but showing no covariance with autosomal recombination rates (Pearson's r = 0.376, P = 0.112 for overall sex-averaged recombination, Supplementary Table 4.2).

In part, the relative weakness of recombination as a predictor may simply have reflected less noise in the estimation of replication time, which has been shown to be conserved between species (Farkash-Amar *et al*., 2008, Chen *et al*., 2010), than in the effective ancestral recombination rate, recombination hotspots known to be fast evolving between even closely related species (Ptak *et al*., 2005, Dumont *et al*., 2011). Nonetheless, the above results suggested that the current focus on recombination associated biased gene conversion as the driver of isochores in mammals may be missing an important contribution from replication timing.

4.3.6 The effect of female recombination on GC has been underestimated owing to interference from replication timing.

The fact that highly recombining domains are GC rich has been taken as evidence that GC rich isochores are structured through gBGC (see Duret and Galtier, 2009 and references therein). Further, it has been suggested that this is a male-driven effect, with GC* covarying more strongly with male than with female recombination rates (Duret and Arndt, 2008, Berglund *et al*., 2009, Dreszer *et al*., 2007, Tyekucheva *et al*., 2008, Webster *et al*., 2005, Galtier *et al*., 2009). Indeed, recently Clément and Arndt (2011) noticed that GC content in rodents was well predicted by male-specific recombination rates but not by female-specific ones. They therefore chose to ignore further consideration of female recombination as a potentially important cause of GC content. The findings presented here raised an interesting possibility: that the gender-specific nature of the impact of gBGC might have been due to the differing relationships between recombination and replication timing in each sex. If it was supposed that some force promotes AT content in late replicating sequence, then if female recombination promotes AT->GC substitutions through biased gene conversion, this unknown force would oppose it. As a consequence, female recombination would leave a diminished footprint of GC->AT biased substitutions than that seen in male meiotic hotspots.

As expected by this model, significant relationships between GC content and female recombination were considerably increased when replication time was controlled for (for the uncontrolled analysis between GC and overall female recombination Spearman's $\rho = 0.038$, $P = 0.025$; controlling for replication time partial Spearman's $\rho = 0.063$, $P = 0.001$, Table 4.1). Indeed, the strength of the correlation, assayed using ρ^2, between GC content and female recombination rates was more than doubled, from 0.00144 to 0.00397, when controlling for replication timing (Table 4.1). By contrast, there was no perceptible change in the relationship between GC and replication time when controlling for any measure of female recombination (for the uncontrolled analysis Spearman's $\rho = 0.293$, $P = 5.34 \times 10^{-71}$; controlling for overall female recombination partial Spearman's $\rho = 0.296$, $P = 0.001$, Table 4.1).

For the influence of male recombination, if anything the covariate uncontrolled analysis would have been expected to over estimate as both early replication timing and higher recombination rates were associated with higher GC content. This was indeed what was observed and again the effect was greatest when the relationship between early replication time and high male recombination rate was strongest: For the uncontrolled analysis between GC and replication time, Spearman's $\rho = 0.293$, $P = 5.34 \times 10^{-71}$; controlling for overall male recombination, partial Spearman's $\rho = 0.292$, $P = 0.001$; controlling for mean-smoothed[2] male recombination, partial Spearman's $\rho = 0.278$, $P = 0.001$ (Table 4.1). Similarly, for the uncontrolled analysis between GC and overall male recombination, Spearman's $\rho = 0.078$, $P = 3.33 \times 10^{-6}$; controlling for replication time, partial Spearman's $\rho = 0.074$, $P = 0.001$ and likewise for the uncontrolled analysis between GC and mean-smoothed[2] male recombination, Spearman's $\rho = 0.144$, $P = 6.96 \times 10^{-18}$; controlling for replication time, partial Spearman's $\rho = 0.109$, $P = 0.001$ (Table 4.1). These effects appear to have been relatively modest corrections, suggesting that the correlation between male recombination rates and local GC content was not grossly misleading.

4.3.7 Why might methodology have influenced the findings?

As the previous analyses showed, using the alternatively curated recombination rate data sets often gave quantitatively different results. In particular, findings obtained from the smoothed versus the gene-focused data sets often qualitatively differed. In

such cases, was one result 'better' than the other and if so, why? In order to answer this question it was necessary to determine both how and why the data sets differed from each other and what these differences suggested about the underlying biological processes.

Despite the use of a number of alternative curation methods, results from the gene-focused data sets did not qualitatively differ from each other (Table 4.1, and Supplementary Tables 4.1 to 4.3). As shown in Figure 4.6 (page 133), this would have been expected given that these methodologies all generated similar, if not the same, recombination rate for each gene. This would imply either that there was little variation in recombination rates at the genic scale or that the marker density was insufficient to detect fine-scale variation in recombination rates. Consistent with the latter explanation, 83% of genes in the final data set did not contain a SNP marker and therefore all gene-focused recombination rates, whether overall or averaged, would have been based on the single recombination rate between the two flanking markers.

In contrast, the two mean-smoothed data sets frequently gave stronger and more highly significant results than did the two median-smoothed data sets. For many of the 2Mb windows across which the smoothing was applied, the distribution of recombination rates was skewed and so statistically, the median would have been the more appropriate measure of centrality to take. However, as Figure 4.7 (page 134) shows, by taking the median recombination rate across each window, as opposed to the mean, genic recombination rates were mostly reduced to zero. By doing so, these data sets implied either that virtually no genes undergo any recombination events or that recombination is a function of a much broader region than genic rates were sampled over. As recombination is known to operate at fine scales (Myers *et al.*, 2006), use of medians therefore failed to capture the recombinational profile of the genome. This would explain why comparisons of other factors to these two median-smoothed data sets often failed to retrieve significant relationships. Overall, these results suggest that scale might have been more important than methodology.

Returning then to the original question, what did the different results from the gene-focused and smoothed data sets reveal about the underlying biological processes? From Figures 4.6 and 4.7 it is clear that the raw recombination rate data was extremely noisy. Assignment of gene-focused recombination rates therefore reflected this level of noise. Further, the measures of recombination used here were derived from only 4 generations (Shifman *et al*., 2006) but recombination hotspots have been shown to evolve rapidly between closely related species (Ptak *et al*., 2005) or even within species (Dumont *et al*., 2011) and homologous blocks in murids show no correlation in recombination rates (Jensen-Seaman *et al*., 2004). As such, while a gene may currently be located in a recombination hotspot, it is unlikely to have been exposed to this recombination rate since speciation. It was therefore unlikely that this noise, reflected in the gene-focused data sets, reflected the ancestral recombination rate of a gene. Given that when a hotspot goes extinct, a new one tends to appear nearby (Myers *et al*., 2005) mean-smoothed recombination rates might therefore have better captured the recombinational history of a given sequence.

Both K_i and GC content are the product of molecular processes occurring at a given site over longer periods of time than the life of a recombination hotspot. Figure 4.8 (page 135) shows that these rates were more consistent along a given chromosome. It was therefore unsurprising that for both genomic features mean-smoothed recombination rates were a stronger covariate than gene-focused recombination rates. Whilst any relationship between timing of replication and recombination was likely to be mechanistic in origin and as such might lead to an expectation of a stronger covariance with gene-focused than smoothed recombination rates, it was clear from Figure 4.8 that replication occurs across distinct Mb sized domains. Again, it was therefore unsurprising that the smoothed recombination rates had stronger relationships with genic replication times than did the gene-focused rates.

Variable			Statistic	Overall genic	Weighted mean	Base pair median	Interval mean	Interval median	Mean smoothed[1]	Median smoothed[1]	Mean smoothed[2]	Median smoothed[2]
X	Y	Z										
K_i	GC	RR_{SA}	ρ	-0.081	-0.081	-0.081	-0.081	-0.081	-0.089	-0.083	-0.092	-0.081
			P	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
		RR_M	ρ	-0.08	-0.079	-0.079	-0.079	-0.079	-0.085	-0.079	-0.088	-0.077
			P	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
		RR_F	ρ	-0.08	-0.08	-0.08	-0.08	-0.08	-0.081	-0.079	-0.083	-0.078
			P	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
	RT	RR_{SA}	ρ	-0.08	-0.08	-0.081	-0.08	-0.08	-0.086	-0.079	-0.087	-0.077
			P	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
		RR_M	ρ	-0.081	-0.081	-0.081	-0.081	-0.081	-0.089	-0.082	-0.09	-0.082
			P	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
		RR_F	ρ	-0.078	-0.078	-0.079	-0.078	-0.079	-0.076	-0.078	-0.076	-0.077
			P	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
	RR_{SA}	-	ρ	0.045	0.041	0.038	0.045	0.043	0.095	0.084	0.1	0.088
			P	0.007	0.015	0.023	0.007	0.01	1.24×10^{-8}	5.14×10^{-7}	2.39×10^{-9}	1.67×10^{-7}
	RR_M	-	ρ	0.015	0.01	0.006	0.013	0.009	0.057	0.054	0.058	0.057
			P	0.379	0.544	0.73	0.454	0.574	0.001	0.001	0.001	0.001
	RR_F	-	ρ	0.044	0.041	0.039	0.044	0.043	0.084	0.071	0.092	0.08
			P	0.009	0.015	0.02	0.008	0.01	5.22×10^{-7}	2.62×10^{-5}	4.25×10^{-8}	1.81×10^{-6}
	RR_{SA}	GC	ρ	0.051	0.047	0.044	0.05	0.049	0.104	0.088	0.111	0.09
			P	0.002	0.003	0.003	0.001	0.005	0.001	0.001	0.001	0.001
	RR_M	GC	ρ	0.021	0.016	0.013	0.019	0.016	0.067	0.055	0.07	0.056
			P	0.114	0.178	0.234	0.137	0.157	0.001	0.003	0.001	0.001
	RR_F	GC	ρ	0.047	0.044	0.043	0.047	0.046	0.087	0.071	0.096	0.08
			P	0.004	0.005	0.005	0.002	0.004	0.001	0.001	0.001	0.001
	RR_{SA}	RT	ρ	0.042	0.039	0.037	0.043	0.041	0.1	0.082	0.105	0.084
			P	0.008	0.01	0.012	0.005	0.014	0.001	0.001	0.001	0.001
	RR_M	RT	ρ	0.017	0.012	0.01	0.015	0.012	0.068	0.055	0.07	0.058
			P	0.162	0.209	0.3	0.197	0.213	0.001	0.001	0.001	0.001
	RR_F	RT	ρ	0.038	0.035	0.035	0.039	0.038	0.08	0.067	0.088	0.076
			P	0.013	0.015	0.022	0.013	0.015	0.001	0.001	0.001	0.001

Table 4.1: Alternative Spearman's correlations for each method used to curate genic recombination rate data where Z = the controlling variable used in partial Spearman's correlations between variables X and Y; K_i = intronic substitution rate between mouse and rat; RT = median replication time for each gene; GC = repeat masked intronic G+C content for each gene; RR_{SA}, RR_M and RR_F = sex-averaged, male and female genic recombination rates respectively.

Variable			Statistic	Overall genic	Weighted mean	Base pair median	Interval mean	Interval median	Mean smoothed[1]	Median smoothed[1]	Mean smoothed[2]	Median smoothed[2]
X	Y	Z										
GC	RT	RR_{SA}	ρ	0.296	0.295	0.294	0.295	0.295	0.29	0.294	0.289	0.294
			P	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
		RR_M	ρ	0.292	0.291	0.29	0.292	0.291	0.283	0.293	0.278	0.293
			P	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
		RR_F	ρ	0.296	0.296	0.295	0.296	0.296	0.295	0.293	0.296	0.293
			P	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
	RR_{SA}	-	ρ	0.067	0.068	0.077	0.067	0.07	0.102	0.048	0.126	0.031
			P	6.44×10^{-5}	4.80×10^{-5}	4.51×10^{-6}	6.46×10^{-5}	2.76×10^{-5}	1.07×10^{-9}	0.004	4.45×10^{-14}	0.064
	RR_M	-	ρ	0.078	0.078	0.085	0.078	0.081	0.111	0.01	0.144	-0.016
			P	3.33×10^{-6}	3.19×10^{-6}	3.85×10^{-7}	3.79×10^{-6}	1.39×10^{-6}	3.39×10^{-11}	0.539	0	0.343
	RR_F	-	ρ	0.038	0.037	0.048	0.035	0.041	0.027	0.005	0.044	-0.007
			P	0.025	0.026	0.005	0.036	0.015	0.104	0.753	0.008	0.692
	RR_{SA}	RT	ρ	0.081	0.08	0.083	0.079	0.08	0.092	0.057	0.116	0.046
			P	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.005
	RR_M	RT	ρ	0.074	0.073	0.074	0.073	0.074	0.079	0.008	0.109	-0.02
			P	0.001	0.001	0.001	0.001	0.001	0.001	0.318	0.001	0.116
	RR_F	RT	ρ	0.063	0.061	0.064	0.059	0.061	0.048	0.021	0.064	0.011
			P	0.001	0.001	0.001	0.001	0.001	0.007	0.104	0.001	0.252
RT	RR_{SA}	-	ρ	-0.034	-0.027	-0.009	-0.03	-0.022	0.048	-0.024	0.051	-0.045
			P	0.041	0.102	0.578	0.074	0.188	0.005	0.148	0.002	0.008
	RR_M	-	ρ	0.025	0.028	0.049	0.026	0.034	0.122	0.01	0.138	0.01
			P	0.135	0.095	0.004	0.122	0.041	2.6×10^{-13}	0.56	1.2×10^{-16}	0.547
	RR_F	-	ρ	-0.076	-0.07	-0.046	-0.073	-0.061	-0.062	-0.051	-0.056	-0.059
			P	6.34×10^{-6}	3.19×10^{-5}	0.006	1.37×10^{-5}	0	0	0.003	0.001	0
	RR_{SA}	GC	ρ	-0.056	-0.05	-0.033	-0.052	-0.045	0.019	-0.04	0.015	-0.056
			P	0.001	0.003	0.02	0.001	0.005	0.14	0.009	0.189	0.001
	RR_M	GC	ρ	0.002	0.005	0.025	0.003	0.011	0.095	0.007	0.102	0.015
			P	0.443	0.376	0.067	0.423	0.255	0.001	0.349	0.001	0.171
	RR_F	GC	ρ	-0.091	-0.084	-0.063	-0.087	-0.076	-0.073	-0.054	-0.072	-0.06
			P	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001

Table 4.1 continued: Alternative Spearman's correlations for each method used to curate genic recombination rate data where Z = the controlling variable used in partial Spearman's correlations between variables X and Y; K_i = intronic substitution rate between mouse and rat; RT = median replication time for each gene; GC = repeat masked intronic G+C content for each gene; RR_{SA}, RR_M and RR_F = sex-averaged, male and female genic recombination rates respectively.

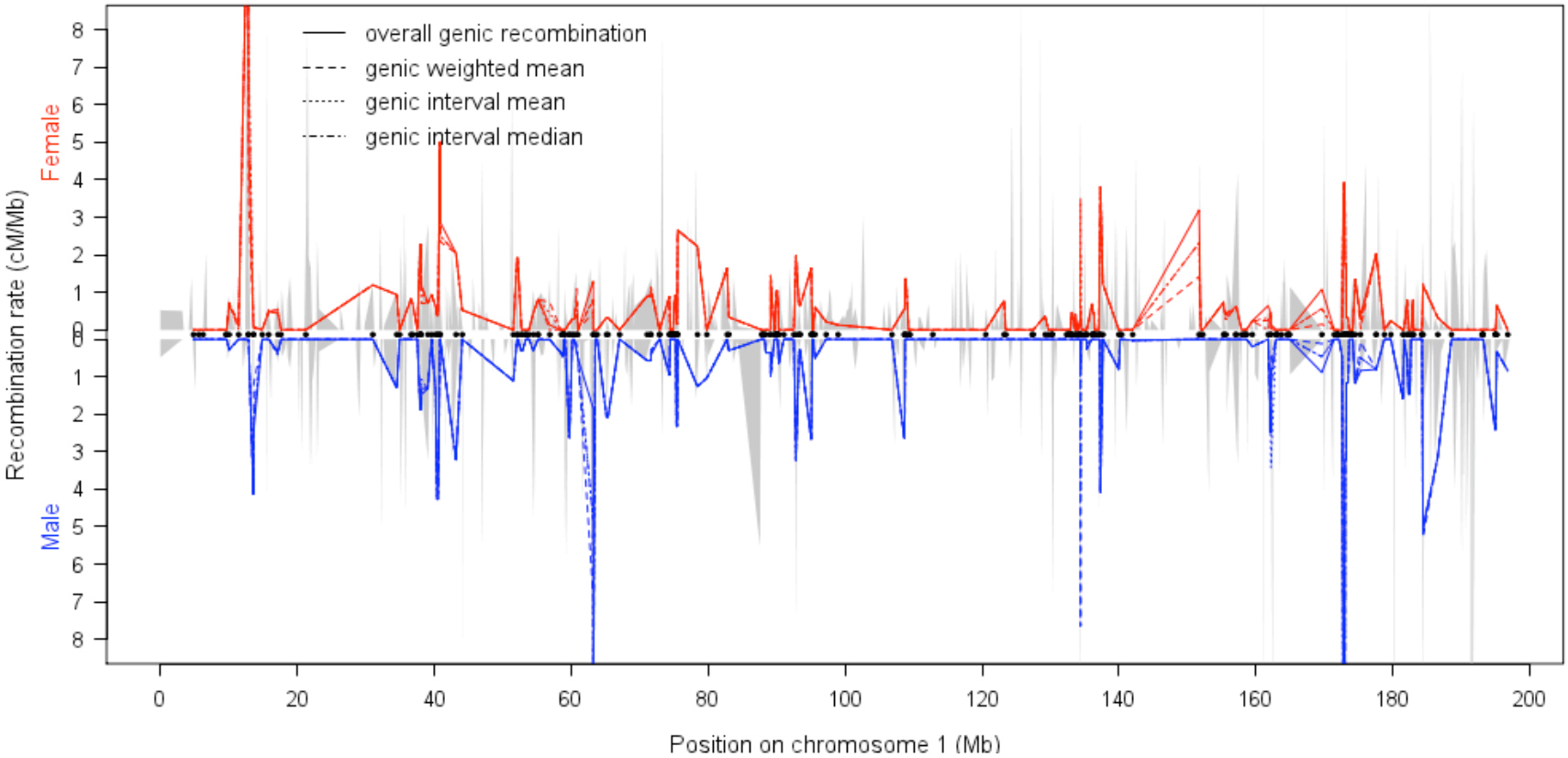


Figure 4.6: Distribution of gene-focused female (red, upper plot) and male (blue, lower plot) recombination rates along mouse chromosome 1. For both genders, the grey shaded plot is the recombination rate between every neighbouring pair of markers. Black dots along the centre of the plot represent genic positions. Lines represent overall (solid), weighted mean (dashed), interval mean (dotted) and interval median (dot/dash) recombination rates assigned to each gene.

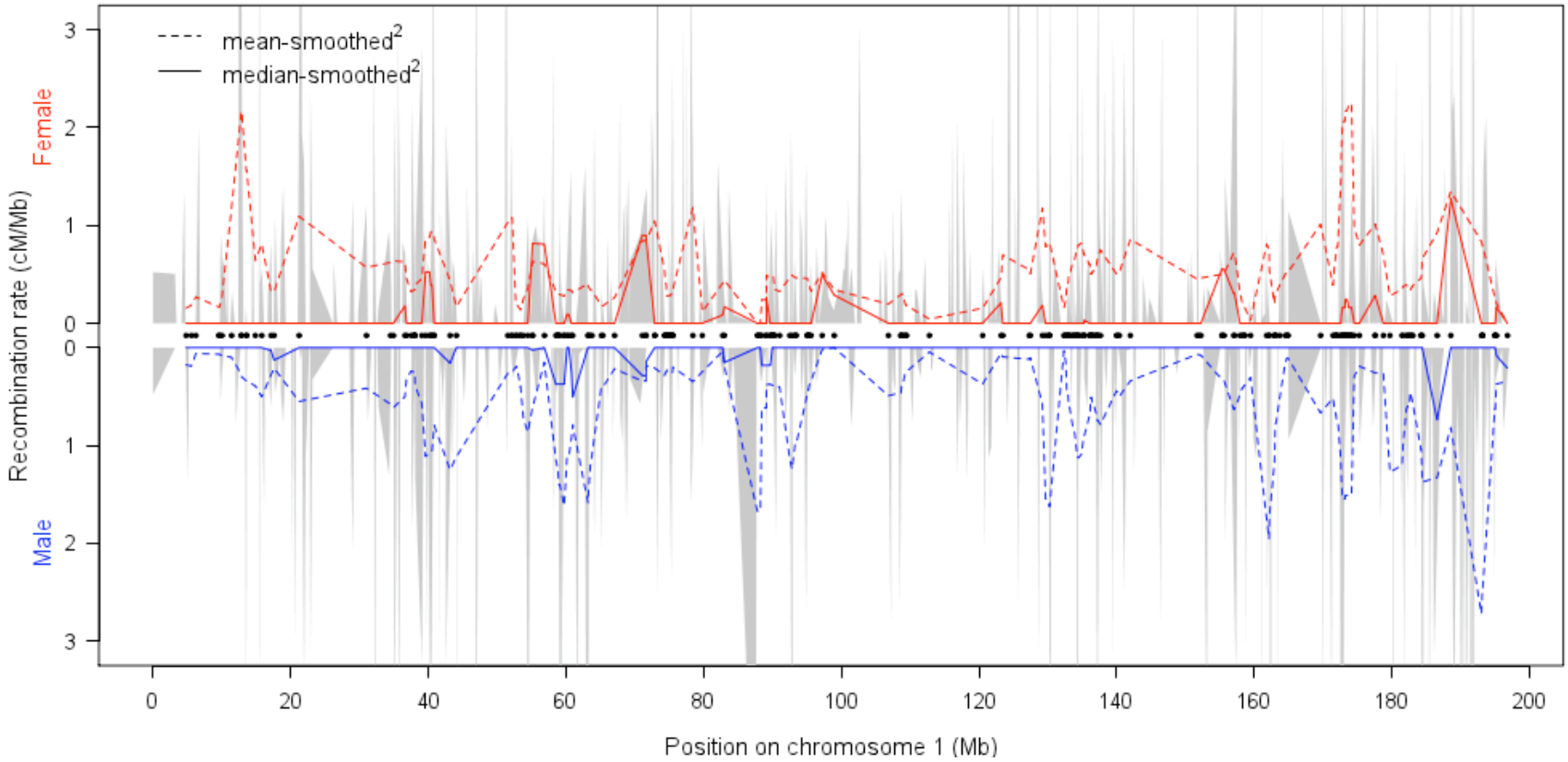


Figure 4.7: Distribution of smoothed female (red, upper plot) and male (blue, lower plot) recombination rates along mouse chromosome 1. For both genders, the grey shaded plot is the recombination rate between every neighbouring pair of markers. Black dots along the centre of the plot represent genic positions. Dotted lines are mean-smoothed^2 genic recombination rates. Solid lines are median-smoothed^2 genic recombination rates.

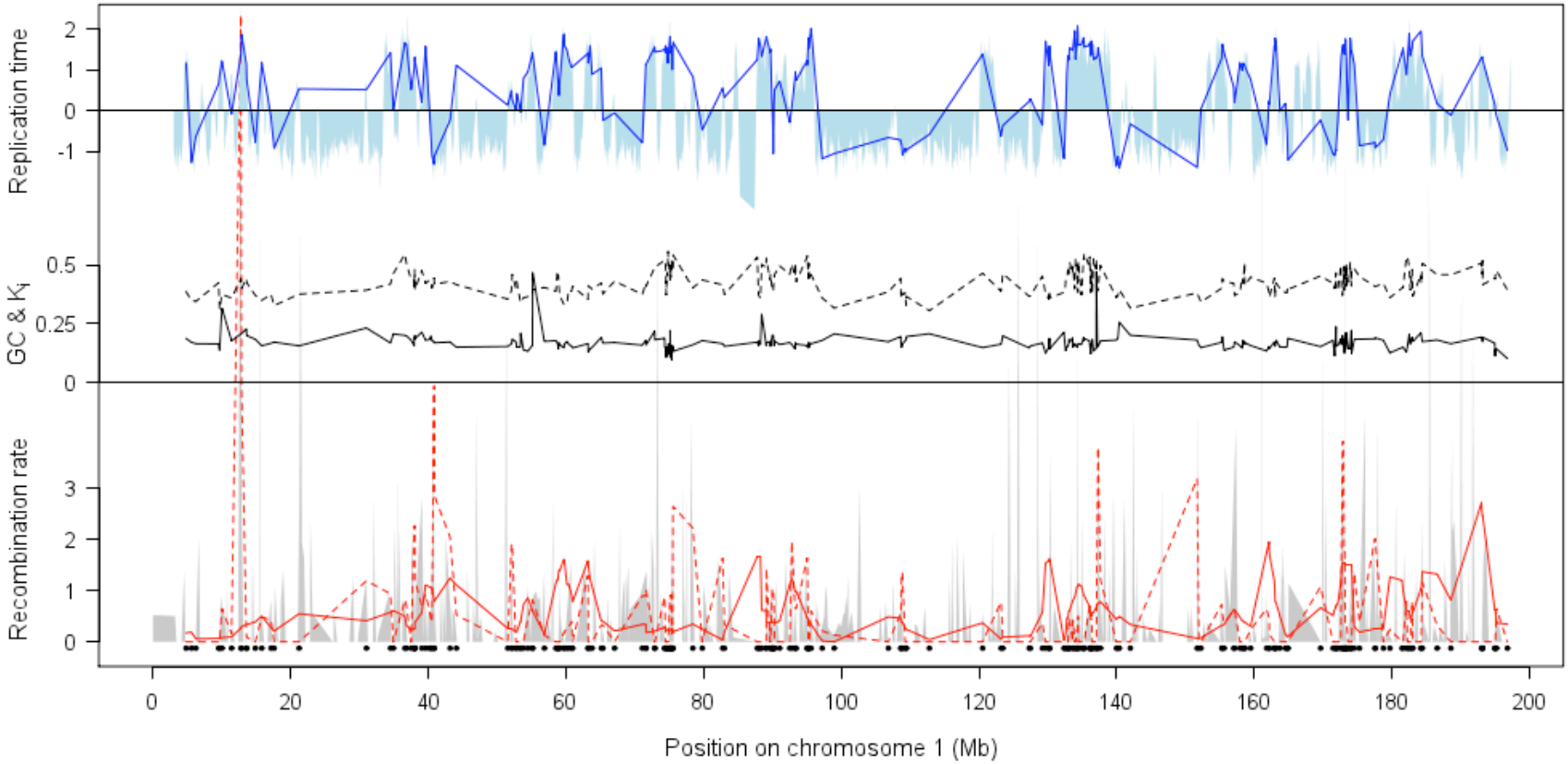


Figure 4.8: Distribution of replication time, GC_i content, K_i and female recombination rates along mouse chromosome 1. Shaded blue area are raw replication times; the solid blue line is the median genic replication time; the black solid line is genic Ki; the black dashed line is repeat masked intronic GC content; the shaded grey area is the recombination rate between every neighbouring pair of markers; the red solid line is the mean-smoothed[2] female recombination rate; the red dashed line is the overall genic female recombination rate. Black dots at the base of the plot represent genic positions.

4.4 Discussion

This chapter was motivated by the hypothesis that, based on previously reported relationships with GC content, early replicating sequences would also be highly recombining. As rates of evolution have been found to be lower where replication is early, but elevated where recombination is higher, it was therefore asked whether the two processes mask each other's impact on neutral substitution rates. What was discovered was that while the use of sex-averaged recombination rates failed to support the initial assumption - that replication time and recombination rate covary - this masked a more important gender-specific complexity that has implications for our understanding of the causes of variation in substitution rate and GC content.

Recent attempts to explain mammalian isochore structure have focused on the role of recombination via the mechanism of GC-biased gene conversion. Evidence for this comes from observations that recombination rate corresponds more strongly to GC* than to current GC, suggesting that recombination is driving GC content (Duret and Arndt, 2008, Meunier and Duret, 2004). In contrast, it is not clear whether GC content determines replication time or *vice versa* and there is evidence for both possibilities (eg. see Chen *et al*., 2010, Hiratani *et al*., 2008). However, the findings presented here suggest that replication time appeared to be as, if not more, important than recombination in relation to GC content.

The idea that the influence of replication time and recombination on GC content may be in opposition is not new. Chen *et al*. (2010) recently reported a greater increase in C:G to A:T substitutions compared to other substitution types as a function of time of replication through S-phase, possibly indicative of a decline in mis-match repair fidelity as replication proceeds. Although these authors noted that the impact of replication timing might therefore counteract the increase in GC arising from gBGC, their use of sex-averaged recombination rates failed to identify that this process is particular to females. The use here of sex-specific data sheds new light on previous observations that gBGC appears to be a male driven phenomena, the impact of female-specific gBGC being countered by later replication forcing higher AT content. This is important as the stronger covariance of GC* with cross-over rates in

males than in females has been taken as evidence against a selectionist explanation for isochore evolution (Duret and Galtier, 2009, Duret and Arndt, 2008). Prior explanations for the origin and maintenance of isochores might therefore warrant re-evaluation with necessary controls for replication timing.

As was shown in Chapter 2 for rat, here a significant increase in intronic rates of evolution where mouse recombination rates are higher was demonstrated. In agreement with estimates in primates (Chen *et al*., 2010), in rodents this was, at most, of about the same magnitude as for replication time, if not weaker. Although it was found that the magnitude of this relationship was overestimated in females and underestimated in males, the corrections were only modest. It was interesting to note that contrary to expectations from primates (Dreszer *et al*., 2007, Berglund *et al*., 2009, Webster *et al*., 2005, Tyekucheva *et al*., 2008, Duret and Arndt, 2008, Galtier *et al*., 2009), the overall relationship between K_i and crossover rates appeared to be driven by recombination in females. This would suggest that the previous model of a male recombination-associated substitution effect to account for elevated and heterogeneous autosomal substitution rates proposed in Chapter 2 might require updating to include an additional or replacement female-specific recombination parameter. Note, however, that allowance for differing male and female recombination effects would not have been possible under the novel implemented in Chapter 2 as there was insufficient information to solve for α, r_m and r_f.

These results suggest that in order to fully understand the relationship between recombination rate and both GC content and substitution rates, it is first necessary to understand how they relate to replication time. Understanding why the relationships differ with respect to gender may be key to this understanding. One possibility may be sexual dimorphism with respect to replication timing. The data used here was derived from male ESC lines but whether these might differ from timings in females is not yet known. As highly expressed genes tend to replicate earlier in S-phase, one might suppose that differences in germline expression might give rise to such sex-specificity in replication time and that this in turn may explain these findings. The possible antagonism between germline expression and recombination (Necsulea *et al*., 2009, McVicker and Green, 2010) suggests the possibility of a unified model in

which differences in germline expression underpin both differences in replication timing and recombination.

All the above results and discussion must by necessity come with the sizeable caveat that the correlations described do not necessarily imply causation. For example, the correlation between GC content and recombination rate might be because a) recombination alters GC content (e.g. via gBGC Duret and Galtier, 2009) b) recombination is more common in GC rich domains (Marsolier-Kergoat and Yeramian, 2009) or c) GC content and recombination covary through a third hidden parameter (possibly gene expression). Indeed, recent attempts to explain mammalian isochore structure have focused on the role of recombination via the mechanism of GC-biased gene conversion (Duret and Galtier, 2009). Evidence for this comes, in part, from observations that recombination rate corresponds more strongly to GC* (predicted equilibrium GC content) than to current GC, suggesting that recombination is driving GC content (Meunier and Duret, 2004, Duret and Arndt, 2008). Experimental evidence (Brown and Jiricny, 1988) that gene conversion, at least in somatic cells, is biased in favour of GC residues over AT ones lends great credence to the model. Further, although GC content and timing of replication were strongly correlated, it is not yet known which is causative of this relationship, nor why, though as previously stated, there is evidence for both possibilities (Chen *et al*., 2010, Hiratani *et al*., 2008). More generally, the strong coupling between isochores and replication timing domains (Watanabe *et al*., 2002, Schmegner *et al*., 2005, Schmegner *et al*., 2007, Woodfine *et al*., 2004, Costantini and Bernardi, 2008) remains both enigmatic and relatively under-explored.

If replication timing is important and causative of isochores then in principle this could be resolved via experimental assays. For example, one hypothesis to explain the high substitution rate in late replicating sequence is that it is caused by error prone translesion synthesis (Lang and Murray, 2011). If translesion synthesis in mammals is biased towards the incorporation of A and T, thereby making late replicating sequence more AT rich, this could then, in principle, explain isochore evolution. This prediction could be examined in mammalian cell lines. Any model suggesting that replication timing causes isochores would also predict that GC rich

sequence, forced by deletion of early and strong replication origins to become late replicating, should start to accumulate A and T.

To conclude, these findings demonstrate the importance of using sex-specific data when investigating drivers of genome evolution such as substitution rates or GC content. As both isochore structure and the process of gBGC is weaker in rodents than primates (Clément and Arndt, 2011), it is recommended that the question of whether replication timing is indeed masking a hitherto unidentified relationship between gBGC and female recombination is subsequently explored in primates.

4.5 References

BERGLUND, J., POLLARD, K. S. & WEBSTER, M. T. (2009) Hotspots of biased nucleotide substitutions in human genes. *PLoS Biol,* 7**,** e1000026.

BROWN, T. C. & JIRICNY, J. (1988) Different base/base mispairs are corrected with different efficiencies and specificities in monkey kidney cells. *Cell,* 54**,** 705-711.

CHAMARY, J.-V. & HURST, L. D. (2004) Similar rates but different modes of sequence evolution in introns and at exonic silent sites in rodents: evidence for selectively driven codon usage. *Mol Biol Evol,* 21**,** 1014-1023.

CHEN, C.-L., RAPPAILLES, A., DUQUENNE, L., HUVET, M., GUILBAUD, G., FARINELLI, L., AUDIT, B., D'AUBENTON-CARAFA, Y., ARNEODO, A., HYRIEN, O. & THERMES, C. (2010) Impact of replication timing on non-CpG and CpG substitution rates in mammalian genomes. *Genome Research,* 20**,** 447-457.

CLÉMENT, Y. & ARNDT, P. F. (2011) Substitution patterns are under different influences in primates and rodents. *Genome Biology and Evolution,* 3**,** 236-245.

COSTANTINI, M. & BERNARDI, G. (2008) Replication timing, chromosomal bands, and isochores. *Proc Natl Acad Sci USA,* 105**,** 3433-3437.

COX, A., ACKERT-BICKNELL, C. L., DUMONT, B. L., DING, Y., BELL, J. T., BROCKMANN, G. A., WERGEDAL, J. E., BULT, C., PAIGEN, B., FLINT, J., TSAIH, S.-W., CHURCHILL, G. A. & BROMAN, K. W. (2009) A new standard genetic map for the laboratory mouse. *Genetics,* 182**,** 1335-1344.

DRESZER, T. R., WALL, G. D., HAUSSLER, D. & POLLARD, K. S. (2007) Biased clustered substitutions in the human genome: the footprints of male-driven biased gene conversion. *Genome Research,* 17**,** 1420-1430.

DUMONT, B. L., WHITE, M. A., STEFFY, B., WILTSHIRE, T. & PAYSEUR, B. A. (2011) Extensive recombination rate variation in the house mouse species complex inferred from genetic linkage maps. *Genome Research,* 21**,** 114-125.

DURET, L. & ARNDT, P. F. (2008) The impact of recombination on nucleotide substitutions in the human genome. *PLoS Genet,* 4**,** e1000071.

DURET, L. & GALTIER, N. (2009) Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu Rev Genomics Hum Genet,* 10**,** 285-311.

DURET, L. & HURST, L. D. (2001) The elevated GC content at exonic third sites is not evidence against neutralist models of isochore evolution. *Mol Biol Evol,* 18**,** 757-762.

FARKASH-AMAR, S., LIPSON, D., POLTEN, A., GOREN, A., HELMSTETTER, C., YAKHINI, Z. & SIMON, I. (2008) Global organization of replication time zones of the mouse genome. *Genome Research,* 18**,** 1562-1570.

GALTIER, N., DURET, L., GLÉMIN, S. & RANWEZ, V. (2009) GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates. *Trends Genet,* 25**,** 1-5.

HELLMANN, I., EBERSBERGER, I., PTAK, S. E., PÄÄBO, S. & PRZEWORSKI, M. (2003) A neutral explanation for the correlation of diversity with recombination rates in humans. *Am J Hum Genet,* 72**,** 1527-1535.

HIRATANI, I., RYBA, T., ITOH, M., YOKOCHI, T., SCHWAIGER, M., CHANG, C.-W., LYOU, Y., TOWNES, T. M., SCHÜBELER, D. & GILBERT, D. M. (2008) Global reorganization of replication domains during embryonic stem cell differentiation. *PLoS Biol,* 6**,** e245.

JENSEN-SEAMAN, M. I., FUREY, T. S., PAYSEUR, B. A., LU, Y., ROSKIN, K. M., CHEN, C.-F., THOMAS, M. A., HAUSSLER, D. & JACOB, H. J. (2004) Comparative recombination rates in the rat, mouse, and human genomes. *Genome Research,* 14**,** 528-538.

KAROLCHIK, D., HINRICHS, A. S., FUREY, T. S., ROSKIN, K. M., SUGNET, C. W., HAUSSLER, D. & KENT, W. J. (2004) The UCSC Table Browser data retrieval tool. *Nucleic Acids Res,* 32**,** D493-D496.

KEIGHTLEY, P. D. & GAFFNEY, D. J. (2003) Functional constraints and frequency of deleterious mutations in noncoding DNA of rodents. *Proc Natl Acad Sci USA,* 100**,** 13402-13406.

LANG, G. I. & MURRAY, A. W. (2011) Mutation rates across budding yeast Chromosome VI are correlated with replication timing. *Genome Biology and Evolution,* 3**,** 799-811.

LERCHER, M. J. & HURST, L. D. (2002) Human SNP variability and mutation rate are higher in regions of high recombination. *Trends Genet,* 18**,** 337-340.

LERCHER, M. J., WILLIAMS, E. J. & HURST, L. D. (2001) Local similarity in evolutionary rates extends over whole chromosomes in human-rodent and mouse-rat comparisons: implications for understanding the mechanistic basis of the male mutation bias. *Mol Biol Evol,* 18**,** 2032-2039.

MAGNI, G. E. (1963) The origin of spontaneous mutations during meiosis. *Proc Natl Acad Sci USA,* 50**,** 975-980.

MAGNI, G. E. & VON BORSTEL, R. C. (1962) Different Rates of Spontaneous Mutation during Mitosis and Meiosis in Yeast. *Genetics,* 47**,** 1097-1108.

MALCOM, C. M., WYCKOFF, G. J. & LAHN, B. T. (2003) Genic mutation rates in mammals: local similarity, chromosomal heterogeneity, and X-versus-autosome disparity. *Mol Biol Evol,* 20**,** 1633-1641.

MARAIS, G. (2003) Biased gene conversion: implications for genome and sex evolution. *Trends Genet,* 19**,** 330-338.

MARSOLIER-KERGOAT, M.-C. & YERAMIAN, E. (2009) GC content and recombination: Reassessing the causal effects for the *Saccharomyces cerevisiae* genome. *Genetics,* 183**,** 31-38.

MCVICKER, G. & GREEN, P. (2010) Genomic signatures of germline gene expression. *Genome Research,* 20**,** 1503-1511.

MEUNIER, J. & DURET, L. (2004) Recombination drives the evolution of GC-content in the human genome. *Mol Biol Evol,* 21**,** 984-990.

MYERS, S., SPENCER, C. C. A., AUTON, A., BOTTOLO, L., FREEMAN, C., DONNELLY, P. & MCVEAN, G. (2006) The distribution and causes of meiotic recombination in the human genome. *Biochem Soc Trans,* 34**,** 526-530.

MYERS, S., BOTTOLO, L., FREEMAN, C., MCVEAN, G. & DONNELLY, P. (2005) A fine-scale map of recombination rates and hotspots across the human genome. *Science,* 310**,** 321-324.

NECSULEA, A., SÉMON, M., DURET, L. & HURST, L. D. (2009) Monoallelic expression and tissue specificity are associated with high crossover rates. *Trends Genet,* 25**,** 519-522.

PERRY, J. & ASHWORTH, A. (1999) Evolutionary rate of a gene affected by chromosomal position. *Curr Biol,* 9**,** 987-989.

PIGANEAU, G., MOUCHIROUD, D., DURET, L. & GAUTIER, C. (2002) Expected relationship between the silent substitution rate and the GC content: implications for the evolution of isochores. *Journal of Molecular Evolution,* 54**,** 129-133.

PINK, C. J. & HURST, L. D. (2010) Timing of replication is a determinant of neutral substitution rates but does not explain slow Y chromosome evolution in rodents. *Mol Biol Evol,* 27**,** 1077-1086.

PINK, C. J., SWAMINATHAN, S. K., DUNHAM, I., ROGERS, J., WARD, A. & HURST, L. D. (2009) Evidence that replication-associated mutation alone does not explain between-chromosome differences in substitution rates. *Genome Biology and Evolution,* 2009**,** 13-22.

PTAK, S., HINDS, D., KOEHLER, K., NICKEL, B., PATIL, N., BALLINGER, D., PRZEWORSKI, M., FRAZER, K. & PÄÄBO, S. (2005) Fine-scale recombination patterns differ between chimpanzees and humans. *Nat Genet,* 37**,** 429-434.

RAMSDELL, C. M., LEWANDOWSKI, A. A., GLENN, J. L. W., VRANA, P. B., O'NEILL, R. J. & DEWEY, M. J. (2008) Comparative genome mapping of the deer mouse (*Peromyscus maniculatus*) reveals greater similarity to rat (*Rattus norvegicus*) than to the lab mouse (*Mus musculus*). *BMC Evol Biol,* 8**,** 65.

SCHMEGNER, C., BERGER, A., VOGEL, W., HAMEISTER, H. & ASSUM, G. (2005) An isochore transition zone in the NF1 gene region is a conserved landmark of chromosome structure and function. *Genomics,* 86**,** 439-445.

SCHMEGNER, C., HAMEISTER, H., VOGEL, W. & ASSUM, G. (2007) Isochores and replication time zones: a perfect match. *Cytogenetic and Genome Research,* 116**,** 167-172.

SHIFMAN, S., BELL, J. T., COPLEY, R. R., TAYLOR, M. S., WILLIAMS, R. W., MOTT, R. & FLINT, J. (2006) A high-resolution single nucleotide polymorphism genetic map of the mouse genome. *PLoS Biol,* 4**,** e395.

STAMATOYANNOPOULOS, J. A., ADZHUBEI, I., THURMAN, R. E., KRYUKOV, G. V., MIRKIN, S. M. & SUNYAEV, S. R. (2009) Human mutation rate associated with DNA replication timing. *Nat Genet,* 41**,** 393-395.

STRATHERN, J. N., SHAFER, B. K. & MCGILL, C. B. (1995) DNA synthesis errors associated with double-strand-break repair. *Genetics,* 140**,** 965-972.

TAMURA, K. & KUMAR, S. (2002) Evolutionary distance estimation under heterogeneous substitution pattern among lineages. *Mol Biol Evol,* 19**,** 1727-1736.

TOUCHON, M., ARNEODO, A., D'AUBENTON-CARAFA, Y. & THERMES, C. (2004) Transcription-coupled and splicing-coupled strand asymmetries in eukaryotic genomes. *Nucleic Acids Res,* 32**,** 4969-4978.

TYEKUCHEVA, S., MAKOVA, K. D., KARRO, J. E., HARDISON, R. C., MILLER, W. & CHIAROMONTE, F. (2008) Human-macaque comparisons illuminate variation in neutral substitution rates. *Genome Biol,* 9**,** R76.

WATANABE, Y., FUJIYAMA, A., ICHIBA, Y., HATTORI, M., YADA, T., SAKAKI, Y. & IKEMURA, T. (2002) Chromosome-wide assessment of replication timing for human chromosomes 11q and 21q: disease-related genes in timing-switch regions. *Hum Mol Genet,* 11**,** 13-21.

WEBSTER, M. T., SMITH, N. G. C., HULTIN-ROSENBERG, L., ARNDT, P. F. & ELLEGREN, H. (2005) Male-driven biased gene conversion governs the evolution of base composition in human alu repeats. *Mol Biol Evol,* 22**,** 1468-1474.

WEDDINGTON, N., STUY, A., HIRATANI, I., RYBA, T., YOKOCHI, T. & GILBERT, D. M. (2008) ReplicationDomain: a visualization tool and comparative database for genome-wide replication timing data. *BMC Bioinformatics,* 9**,** 530.

WOODFINE, K., FIEGLER, H., BEARE, D. M., COLLINS, J. E., MCCANN, O. T., YOUNG, B. D., DEBERNARDI, S., MOTT, R., DUNHAM, I. & CARTER, N. P. (2004) Replication timing of the human genome. *Hum Mol Genet,* 13**,** 191-202.

YAFFE, E., FARKASH-AMAR, S., POLTEN, A., YAKHINI, Z., TANAY, A. & SIMON, I. (2010) Comparative analysis of DNA replication timing reveals conserved large-scale chromosomal architecture. *PLoS Genet,* 6**,** e1001011.

Chapter 5. **Discussion**

This thesis has presented evidence that the number of DNA replications is not the sole determinant of mutation rate variability across rodent genomes. In particular, it was shown that, contrary to the predictions of the theory of male driven evolution, the Y chromosome was not the fastest evolving chromosomal type. Novel models that incorporated an additional recombination parameter were proposed to account for the elevated autosomal rate of evolution. In violation of the assumption that replication-associated errors occur randomly across the genome, mutation rates were shown to be elevated in genes that replicate later during S-phase. A previously unidentified sex-specific covariance between replication timing and recombination rate was shown that resulted in a moderate underestimation of the relationship between recombination rate and divergence in males and a slight overestimation of this relationship in females. Similarly, it was shown that although recombination and early replication both act in the same direction with respect to GC content in males, the opposite is true in females. In the latter, later replication possibly acts to increase AT content, but the high recombination rates experienced by the same sequences counters this increase by elevating GC content. Finally, significant inter-autosomal variability in rodent divergence was confirmed, with highly rearranged autosomes tending to evolve faster. Although replication timing was able to explain some of the autosomal variation in rates, it could not account for why the late replicating Y-chromosome did not evolve faster.

The unexpectedly slow evolving Y-chromosome was an interesting finding, but one that requires confirmation both in other types of sequence and in different groups. If true, then explanations for this low rate of evolution are required. This thesis proposed that Y-linked sequence is not exposed to the recombination-associated substitutions, which autosomal and X-linked sequences accumulate. Alternatively, it might be expected that selection would favour enhanced repair in the male germline to deal with the higher mutational input. There is some evidence that proteins involved in the repair of double strand breaks localise to the XY body (Handel, 2004) and that in the zygote homologous recombination is more active in the paternal pronucleus (Derijck *et al*., 2008). However, whether these directly repair

replication errors incurred during maintenance of spermatogonia is not known. Further research into the specific nature and efficiency of repair in each germ-line would be extremely informative. Secondly, why the impact of later replication on divergence does not appear to extend to Y-linked sequence is unclear. Finally, a rapidly evolving Y chromosome is generally considered evidence of a higher mutational input from older fathers. If this is not the case then the prevailing theory, which has public health implications, needs revising.

The novel models proposed in this thesis require further investigation and the Drosophilids represent an ideal group in which to do this. As flies have a short generation time, the expectation under Miyata *et al*.'s (1987) original model would be that neutral sequence on all types of chromosome should evolve at the same rate, giving $\alpha = 1$ (Bauer and Aquadro, 1997). This should make it relatively simple to identify the impact of recombination on chromosomal substitution rates. As male *Drosophila* are achiasmate, the first novel model (Equations 10 and 11, Section 2.3.5) that incorporates a general recombination-associated substitution effect would not differ from a model that incorporated only a female-specific recombination parameter. Under this model, the X-chromosome would be exposed to recombination-associated substitutions two thirds of the time, compared to the autosomes that would only be exposed to this effect half of the time. The non-recombining Y chromosome would therefore evolve slowest. In contrast, if only male recombination is important, then the second novel model (Equation 14, Section 2.3.5) would not differ from Miyata *et al*.'s (1987) model (Equation 2, Section 2.1), since the recombination effect would be equal to zero. Here then, the different chromosomal types should evolve at the same rate, as none would receive a recombination-associated boost to the substitution rate. An additional advantage of this work would be the opportunity to re-examine a report suggesting that despite the short generation time of flies, $\alpha = 2$ (Bachtrog, 2008). There are concerns over the methodologies employed in this study that failed to control for ancestral polymorphisms, hence the desire to repeat the work using a more accurate approach. Whilst care must also be taken to avoid sites potentially under selection, thus restricting the analysis to exon cores (Warnecke and Hurst, 2007, Warnecke *et al*., 2008) and short introns (Halligan and Keightley, 2006, Parsch *et al*., 2010), the

availability of replication timing data (see Appendix 1) would enable substitutions arising from this (Weber *et al*., 2012) to be controlled for.

Aside from the issue of the slow evolving Y-chromosome, this thesis showed that Miyata *et al*.'s (1987) model could not be used to estimate the extent of male bias in the mutation rate. One possible explanation for this is that the model may have been incorrectly applied. Although Miyata *et al*.'s (1987) model assumed that the mutation rate was accurately measured, this might not have been the case. It is now well established that synonymous sites do not evolve neutrally, owing to the need to specify mRNA secondary structure (Chamary and Hurst, 2005, Stoletzki, 2008), preserve splice enhancer or suppressor sites (Parmley *et al*., 2006), ensure accurate translation (Parmley and Huynen, 2009, Zhou *et al*., 2009) or possibly owing to use of preferred codons (Chamary and Hurst, 2004, Waldman *et al*., 2011). However, for this to explain the discrepancy from the expected results using the exonic data set, synonymous sites on the Y chromosome would have to be under stronger or more efficient purifying selection than those on the autosomes. One possible difference is that new mutations on an autosome are less likely to be phenotypically exposed than those on the haploid Y chromosome. By contrast, as it has a low effective population size, weakly deleterious mutations should be fixed on the Y chromosome at a higher rate, all else being equal. Further, purifying selection on synonymous sites does not explain why the same result was retrieved from the intronic data set, especially as filters were applied to remove any sites potentially under selective constraints. Despite this, it is possible that some introns containing functional sequence might have escaped these filters so that substitution rates were derived from sites that were not evolving neutrally. It is also important to consider that if gBGC does indeed impact on substitution rates, then there are potentially no sites not subject to selection or fixation biases in the genome. This would prevent the use of molecular divergence as an accurate measurement of the mutation rate.

A further issue relating to the application of Miyata *et al*.'s (1987) model is that, due to concerns over the ability to accurately assign CpG dinucleotides from pairwise alignments (Gaffney and Keightley, 2008) the analyses presented in this thesis were based on all sites rather than evaluating substitution rates at CpG and non-CpG sites

separately. Mutations at CpGs occur independently of replication and as such, are not expected to demonstrate a male effect. Indeed, Taylor *et al*. (2006) provide evidence consistent with this in primates with $\alpha_{XAutosome} \sim 6$ at non-CpG sites reduced to $\alpha_{XAutosome} \sim 2$ at CpG sites, although assignment of CpGs was subject to the same estimation error using pairwise alignments described by Gaffney and Keightley (2008). The inclusion of CpG sites in the analyses presented in this thesis might account for some of the discrepancy in estimates of α: α estimated from each of the three pairwise comparisons appeared also to be discrepant in the human-chimpanzee comparison when all sites were used, but removal of CpG sites resulted in some degree of convergence (Taylor *et al*., 2006, supplementary information). Note, however, that the implications of these results for the accuracy of Miyata *et al*.'s model were not recognised and so significance of the effect was not tested.

It is perhaps also worth noting here that the negative relationship between K_i and GC_i identified in this thesis adds to the debate about the nature of the relationship between GC content and divergence. Previous observations of a positive relationship have generally been derived from exonic four-fold degenerate sites (Hurst and Williams, 2000). There are at least two reasons to suppose that the data from introns might be more reliable. First, third sites are potentially under selective constraints, as discussed (but see Duret and Hurst, 2001). Second, and possibly more importantly, owing to the structure of the genetic code, three quarters of 4-fold degenerate codons have a C or G at the second site (G or C at second site: TCN, CGN, GCN, GGN, ACN, CCN; A or T at second site: GTN, CTN). Consequently the third site dinucleotide context of four-fold degenerate codons is far from random, causing possibly higher rates of CpG mutation when third site GC content is high. Whether the relationship between K_i and GC_i is indeed linear, as assumed here, or follows a curved distribution, is left to future work.

Even if correctly applied, Miyata *et al*.'s (1987) model is still likely to be flawed, assuming as it does one dominant source of new mutations. As explored throughout this thesis, a number of other determinants of substitution rates have been identified. In particular it was shown that intronic divergence could be explained by differences in replication timing across the rodent genome, providing strong evidence that

replication associated mutations do not occur randomly across the genome per replication event, as the model inherently assumes. Why early replicating sequence evolves slower than those which replicate later is not yet fully understood, although in yeast, there is compelling evidence for one hypothesis (Lang and Murray, 2011). Unrepaired lesions persisting from G1 phase impede progress of replicative polymerases, resulting in a de-coupling of leading and lagging strand synthesis and the formation of single stranded gaps. Early in S-phase, the largely non-mutagenic method of template switching is employed to fill in these gaps. However, in a 'last ditch' attempt to repair single stranded DNA late in S-phase, slow and error prone translesion polymerases are used. Consistent with this hypothesis, deletion of Rev1 in yeast, a translesion polymerase not expressed until late S-phase (Waters and Walker, 2006), resulted in disruption of translesion synthesis and nearly a five-fold reduction in the mutation rate of fast evolving late replicating sequence (Lang and Murray, 2011). It would be valuable to know both whether the same is true in mammalian cells and also whether other repair systems show temporal changes across the cell cycle, with error prone mechanisms more active in late S and G2 phase. Whatever the mechanism, the trend for late replication sequence to have high neutral substitution rates appears to be taxonomically common (Stamatoyannopoulos *et al*., 2009, Chen *et al*., 2010, Weber *et al*., 2012, Flynn *et al*., 2010, Lang and Murray, 2011).

All studies appear to be in general agreement about the approximate magnitude of the difference between early and late replicating sequence. The 10.5% increase in rodent divergence as replication timing proceeds over S-phase is on a par with that recovered in rodents from a subsequent study (16%) (Chen *et al*., 2010). Why the difference in primates (22-28%) (Stamatoyannopoulos *et al*., 2009, Chen *et al*., 2010) is a little greater is not clear. The difference in Drosophilids is in agreement with that seen in rodents (~10% increase in diversity, intronic divergence and dS). That the latter rises to ~30% when removing genes subject to strong codon usage bias (Weber *et al*., 2012) suggests that the lower estimates may well be effected by the influence of selection. Indeed, in genes with very strong codon bias, the difference diminishes to well under 10% and becomes non-significant (Weber *et al*., 2012).

Alternatively Chen *et al*. (2010) suggest the low increase in rodent divergence may stem from substitution saturation as they find a 30% increase in mouse diversity in later S-phase. There may also be a statistical element to the discrepancy with Stamatoyannopoulos *et al*.'s (2009) study as this was based on four large bins (see Sémon *et al*., 2005). To date, none of the studies examining this issue in mammals have used coding sequence, so future work is required to determine if the effect is stronger at synonymous sites, as observed in *Drosophila* species, or whether the latter is owing to intronic sites in these species being subject to stronger purifying selection than core synonymous sites.

That early replication has been associated with low rates of evolution and, to some extent, gene expression, might have further implications. It has been observed (Chuang and Li, 2004) that genes involved in essential processes tend to be located in genomic regions with low mutation rates, whereas those involved in extracellular processes tend to evolve faster at four-fold degenerate sites. This has been interpreted either as selection for gene location with respect to regional variations in the mutation rate or local adaptation of the mutation rate depending on whether new mutations might be deleterious or advantageous (Chuang and Li, 2004). However, the strength of selection acting on the local mutation rate or on a relocation event must be incredibly weak, as the mutation rate is extremely low and the absolute difference between hot and cold spots is relatively modest (see Hodgkinson and Eyre-Walker, 2011). Whether these gene categories tend to be early or late replicating and whether this might instead explain their rates of evolution would present an interesting avenue of future research.

That late replicating domains tend to have high recombination rates in females but low recombination rates in males and the impact that this relationship has both on divergence and GC content is important, in particular as it provides a potential explanation as to why male, but not female recombination rates have previously been found to covary with clusters of substitutions associated with gBGC (e.g. Berglund *et al*., 2009, Dreszer *et al*., 2007). However, this work leaves unresolved the major question as to why this relationship differs between the two genders. As previously

discussed, differences in chromatin structure have been suggested as an underlying cause of gender differences in recombination (Petkov *et al*., 2007). Further, elevated rates of recombination have been observed in imprinted and randomly monoallelic genes compared to those with biallelic expression, suggestive of an adverse relationship between transcription and recombination (Necsulea *et al*., 2009). Similarly, chromatin structure and expression are strong, though not perfect, covariates of replication timing (see Farkash-Amar and Simon, 2009 and references therein).

It is therefore possible to speculate that differences in expression and chromatin structure between the two germ-lines might explain the observations presented in this thesis. Consistent with this hypothesis, a recent report identified that in meiotic tissues, high gene expression was related to low rates of crossover. Further, the strength of this relationship appeared to be somewhat stronger in tissues derived from the female than the male germline (McVicker and Green, 2010). Future work exploring this issue would benefit from additional replication timing data determined both from males and from germline cells. However, it is likely to be complicated by difficulties in obtaining relevant expression data. This is a complex issue as most germline expression sets are derived from terminally differentiated germ cells and as such may contain transcripts from post-meiotic expression (Vibranovski *et al*., 2010), be biased in favour of transcripts that have been stored in the cytoplasm (Schäfer *et al*., 1995), or contain transcripts not directly expressed by the gamete, but passed to it from nurse cells (Pepling and Spradling, 1998).

Another unexplained observation presented in this thesis is that highly rearranged mouse autosomes tend to evolve faster at intronic sites. It is interesting to note that a similar trend has been observed in mitochondrial genomes (Xu *et al*., 2006, Shao *et al*., 2003). Why this might be is unknown and therefore requires further investigation. Although divergence in replication timing has been observed at fusion breakpoints (Yaffe *et al*., 2010), it is unlikely that this would explain the observed relationship. Fusions tend to occur between domains of similarly timed replication, these tending to physically interact in the nucleus. In the rare scenarios where early- and late-replicating domains fuse, early replication invades the late domain, thus

advancing the latter's replication time and subsequently slowing its rate of evolution. However, it is interesting to note that rearrangements, perhaps unsurprisingly, tend to occur between closely interacting regions of the genome occupying the same nuclear compartment (e.g. Yaffe *et al*., 2010, Wijchers and de Laat, 2011). As more is learnt about the spatial distribution within the nucleus of both the genome and of factors such as repair machinery that associate with it, it will be interesting to see how this impacts on the evolution of other genomic features. For example, replication timing may be a proxy for nuclear location.

With the benefit of hindsight, it is worth asking whether the analyses presented in this thesis could have been improved. For example, were rodents the best group to examine these issues in? The availability of additional mouse Y-linked sequence enabled a rigorous re-testing of Miyata's model over that performed previously (Smith and Hurst, 1999). This, together with the opportunity to determine whether the relationship between replication timing and evolutionary rates extended beyond primates, would defend the choice of rodents. In contrast, lengthening of the chromosomal arms, in part due to relocation of the centromere from a metacentric to acrocentric position in rodents, has resulted in a reduced rate of meiotic recombination. This is thought to be responsible for a decline in isochore structure in rodents, the so-called 'minor shift' (Clément and Arndt, 2011, but see Duret *et al*., 2006). Moreover, although gBGC has been shown to operate in rodents, its impact on GC content is weaker than in primates (Clément and Arndt, 2011). Therefore, rodents may not have been the ideal species in which to explore the interaction between replication timing and recombination rates with regard to GC content. Future work in species such as primates, which have a stronger isochore structure and more prominent gBGC, may find stronger effects.

5.3 References

BACHTROG, D. (2008) Evidence for male-driven evolution in *Drosophila*. *Mol Biol Evol,* 25, 617-619.

BAUER, V. L. & AQUADRO, C. F. (1997) Rates of DNA sequence evolution are not sex-biased in *Drosophila melanogaster* and *D. simulans*. *Mol Biol Evol,* 14, 1252-1257.

BERGLUND, J., POLLARD, K. S. & WEBSTER, M. T. (2009) Hotspots of biased nucleotide substitutions in human genes. *PLoS Biol,* 7, e1000026.

CHAMARY, J.-V. & HURST, L. D. (2004) Similar rates but different modes of sequence evolution in introns and at exonic silent sites in rodents: evidence for selectively driven codon usage. *Mol Biol Evol,* 21, 1014-1023.

CHAMARY, J. V. & HURST, L. D. (2005) Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. *Genome Biol,* 6, R75.

CHEN, C.-L., RAPPAILLES, A., DUQUENNE, L., HUVET, M., GUILBAUD, G., FARINELLI, L., AUDIT, B., D'AUBENTON-CARAFA, Y., ARNEODO, A., HYRIEN, O. & THERMES, C. (2010) Impact of replication timing on non-CpG and CpG substitution rates in mammalian genomes. *Genome Research,* 20, 447-457.

CHUANG, J. H. & LI, H. (2004) Functional bias and spatial organization of genes in mutational hot and cold regions in the human genome. *PLoS Biol,* 2, e29.

CLÉMENT, Y. & ARNDT, P. F. (2011) Substitution patterns are under different influences in primates and rodents. *Genome Biology and Evolution,* 3, 236-245.

DERIJCK, A., VAN DER HEIJDEN, G., GIELE, M., PHILIPPENS, M. & DE BOER, P. (2008) DNA double-strand break repair in parental chromatin of mouse zygotes, the first cell cycle as an origin of de novo mutation. *Hum Mol Genet,* 17, 1922-1937.

DRESZER, T. R., WALL, G. D., HAUSSLER, D. & POLLARD, K. S. (2007) Biased clustered substitutions in the human genome: the footprints of male-driven biased gene conversion. *Genome Research,* 17, 1420-1430.

DURET, L., EYRE-WALKER, A. & GALTIER, N. (2006) A new perspective on isochore evolution. *Gene,* 385, 71-74.

DURET, L. & HURST, L. D. (2001) The elevated GC content at exonic third sites is not evidence against neutralist models of isochore evolution. *Mol Biol Evol,* 18, 757-762.

FARKASH-AMAR, S. & SIMON, I. (2009) Genome-wide analysis of the replication program in mammals. *Chromosome Res,* 18, 115-125.

FLYNN, K. M., VOHR, S. H., HATCHER, P. J. & COOPER, V. S. (2010) Evolutionary rates and gene dispensability associate with replication timing in the archaeon *Sulfolobus islandicus*. *Genome Biology and Evolution,* 2, 859-869.

GAFFNEY, D. J. & KEIGHTLEY, P. D. (2008) Effect of the assignment of ancestral CpG state on the estimation of nucleotide substitution rates in mammals. *BMC Evol Biol,* 8, 265.

HALLIGAN, D. L. & KEIGHTLEY, P. D. (2006) Ubiquitous selective constraints in the *Drosophila* genome revealed by a genome-wide interspecies comparison. *Genome Research,* 16, 875-884.

HANDEL, M. A. (2004) The XY body: a specialized meiotic chromatin domain. *Exp Cell Res,* 296, 57-63.

HODGKINSON, A. & EYRE-WALKER, A. (2011) Variation in the mutation rate across mammalian genomes. *Nat Rev Genet,* 12, 756-766.

HURST, L. D. & WILLIAMS, E. J. (2000) Covariation of GC content and the silent site substitution rate in rodents: implications for methodology and for the evolution of isochores. *Gene,* 261, 107-114.

LANG, G. I. & MURRAY, A. W. (2011) Mutation rates across budding yeast Chromosome VI are correlated with replication timing. *Genome Biology and Evolution,* 3, 799-811.

MCVICKER, G. & GREEN, P. (2010) Genomic signatures of germline gene expression. *Genome Research,* 20, 1503-1511.

MIYATA, T., HAYASHIDA, H., KUMA, K. & YASUNAGA, T. (1987) Male-driven molecular evolution demonstrated by different rates of silent substitutions between autosome-and sex chromosome-linked genes. *Proceedings of the Japan Academy. Ser. B: Physical and Biological Sciences,* 63, 327-331.

NECSULEA, A., SÉMON, M., DURET, L. & HURST, L. D. (2009) Monoallelic expression and tissue specificity are associated with high crossover rates. *Trends Genet,* 25, 519-522.

PARMLEY, J. L., CHAMARY, J. V. & HURST, L. D. (2006) Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. *Mol Biol Evol,* 23, 301-309.

PARMLEY, J. L. & HUYNEN, M. A. (2009) Clustering of codons with rare cognate tRNAs in human genes suggests an extra level of expression regulation. *PLoS Genet,* 5, e1000548.

PARSCH, J., NOVOZHILOV, S., SAMINADIN-PETER, S. S., WONG, K. M. & ANDOLFATTO, P. (2010) On the utility of short intron sequences as a reference for the detection of positive and negative selection in *Drosophila*. *Mol Biol Evol,* 27, 1226-1234.

PEPLING, M. E. & SPRADLING, A. C. (1998) Female mouse germ cells form synchronously dividing cysts. *Development,* 125, 3323-3328.

PETKOV, P. M., BROMAN, K. W., SZATKIEWICZ, J. P. & PAIGEN, K. (2007) Crossover interference underlies sex differences in recombination rates. *Trends in Genetics,* 23, 539-542.

SCHÄFER, M., NAYERNIA, K., ENGEL, W. & SCHÄFER, U. (1995) Translational control in spermatogenesis. *Dev Biol,* 172, 344-352.

SÉMON, M., MOUCHIROUD, D. & DURET, L. (2005) Relationship between gene expression and GC-content in mammals: statistical significance and biological relevance. *Hum Mol Genet,* 14, 421-427.

SHAO, R., DOWTON, M., MURRELL, A. & BARKER, S. C. (2003) Rates of gene rearrangement and nucleotide substitution are correlated in the mitochondrial genomes of insects. *Mol Biol Evol,* 20, 1612-1619.

SMITH, N. G. & HURST, L. D. (1999) The causes of synonymous rate variation in the rodent genome. Can substitution rates be used to estimate the sex bias in mutation rate? *Genetics,* 152, 661-673.

STAMATOYANNOPOULOS, J. A., ADZHUBEI, I., THURMAN, R. E., KRYUKOV, G. V., MIRKIN, S. M. & SUNYAEV, S. R. (2009) Human mutation rate associated with DNA replication timing. *Nat Genet,* 41, 393-395.

STOLETZKI, N. (2008) Conflicting selection pressures on synonymous codon use in yeast suggest selection on mRNA secondary structures. *BMC Evol Biol,* 8, 224.

TAYLOR, J., TYEKUCHEVA, S., ZODY, M., CHIAROMONTE, F. & MAKOVA, K. D. (2006) Strong and weak male mutation bias at different sites in the primate genomes: insights from the human-chimpanzee comparison. *Mol Biol Evol,* 23, 565-573.

VIBRANOVSKI, M. D., CHALOPIN, D. S., LOPES, H. F., LONG, M. & KARR, T. L. (2010) Direct evidence for postmeiotic transcription during *Drosophila melanogaster* spermatogenesis. *Genetics,* 186, 431-433.

WALDMAN, Y. Y., TULLER, T., KEINAN, A. & RUPPIN, E. (2011) Selection for translation efficiency on synonymous polymorphisms in recent human evolution. *Genome Biology and Evolution,* 3, 749-761.

WARNECKE, T. & HURST, L. D. (2007) Evidence for a trade-off between translational efficiency and splicing regulation in determining synonymous codon usage in *Drosophila melanogaster*. *Mol Biol Evol,* 24, 2755-2762.

WARNECKE, T., PARMLEY, J. L. & HURST, L. D. (2008) Finding exonic islands in a sea of non-coding sequence: splicing related constraints on protein composition and evolution are common in intron-rich genomes. *Genome Biol,* 9, R29.

WATERS, L. S. & WALKER, G. C. (2006) The critical mutagenic translesion DNA polymerase Rev1 is highly expressed during G_2/M phase rather than S phase. *Proc Natl Acad Sci USA,* 103, 8971-8976.

WEBER, C. C., PINK, C. J. & HURST, L. D. (2012) Late-Replicating Domains Have Higher Divergence and Diversity in *Drosophila melanogaster*. *Molecular Biology and Evolution*. 29, 873–882.

WIJCHERS, P. J. & DE LAAT, W. (2011) Genome organization influences partner selection for chromosomal rearrangements. *Trends Genet,* 27, 63-71.

XU, W., JAMESON, D., TANG, B. & HIGGS, P. G. (2006) The relationship between the rate of molecular evolution and the rate of genome rearrangement in animal mitochondrial genomes. *Journal of Molecular Evolution,* 63, 375-392.

YAFFE, E., FARKASH-AMAR, S., POLTEN, A., YAKHINI, Z., TANAY, A. & SIMON, I. (2010) Comparative analysis of DNA replication timing reveals conserved large-scale chromosomal architecture. *PLoS Genet,* 6, e1001011.

ZHOU, T., WEEMS, M. & WILKE, C. O. (2009) Translationally optimal codons associate with structurally sensitive sites in proteins. *Mol Biol Evol,* 26, 1571-1580.

Appendix 1. **A Tale of Two Data Sets: Curation of *Drosophila melanogaster* Replication Times**

Unpublished. Data set used for analyses published as:

Late replicating domains have higher divergence and diversity in *Drosophila melanogaster*.

Claudia C. Weber, Catherine J. Pink, and Laurence D. Hurst

Molecular Biology and Evolution (2012). 29(2), 873-882

As described in Chapters 3 and 4, earlier replicating sequences have been shown to have lower substitution rates both in primates (Stamatoyannopoulos *et al*., 2009, Chen *et al*., 2010) and in rodents (Pink and Hurst, 2010), but the same had not yet been demonstrated in any non-mammalian metazoans. As flies show variation in replication timing (MacAlpine *et al*., 2004), they represented an ideal species in which to address this issue. The impact of replication timing on both synonymous and intronic divergence plus diversity was therefore analysed in Drosophilids (Weber *et al*., 2012). This analysis required curation of a data set of replication timing values for *Drosophila melanogaster*. The methods used to generate this data set are presented here and demonstrate the importance of careful handing of data obtained from publicly accessible repositories.

Replication times in *D.melanogaster* were measured by Dirk Schübeler's group at the Friedrich Miescher Institute for Biomedical Research using the same methods as David Gilbert's group (e.g. see Gilbert, 2010) applied to mouse, which were described in Chapter 1. Again, this generated a data set of $\log_2$ ratios of early to late replicating fractions, where positive values were indicative of earlier replication and negative values indicate later replication. Array probes were located at 35 bp intervals along the genome (Schwaiger *et al*., 2009). Of the cell types for which replication timing data was measured, embryonic derived Kc cells were considered to be the closest cell type to the germline and were therefore chosen for the analysis.

Schwaiger *et al*. made their data available in two locations: The NCBI Gene Expression Omnibus (GEO), series GSE13328, data set GSE336362, file GSM336362_Kc_replication_timing.txt (hereafter called the GEO data set); and the Replication Domain database, file RD_KcRT_Schwaiger2009.txt (hereafter called the Replication Domain data set).

Although putatively the same data set, it was immediately apparent that the two sources differed in the formatting of the data, notably the number and naming of columns and the number of rows. This prompted a more detailed examination of each data set, which revealed that whilst all 3,159,411 probes in the GEO data set were found to occur uniquely, of the 3,159,096 probes in the Replication Domain data set, 24,662 were found to be duplicates, whereby two or more array probes were found to have the same chromosome and start positions. Of these duplicates, 20,189 probes had the same replication timing value, but 4,473 differed in their replication times. To determine why the two putatively identical data sets differed in this way, the authors of both sources were contacted.

Both Dr Schwaiger, who deposited the original data on GEO, and Dr Hiratani of Replication Domain explained what the data in the columns of the two data sets represented. Further, Dr Hiratani kindly supplied copies of all emails pertaining to the publication of the data on the Replication Domain website. He was also able to account for the different sample sizes of each data set, explaining that he had removed 315 probes that did not have a chromosomal location, leaving 3,159,096 rows. However, although Dr Schwaiger advised that GEO data set was based on assembly dm3, Dr Hiratani believed that the data had only been lifted to assembly dm3 for publication on the Replication Domain database, and that the GEO data set might therefore be based on assembly dm2. Suggestions made by the authors that might account for the duplicates in the Replication Domain data set included either that the curators of the Replication Domain database may have averaged replication times over genes or windows, or that the original authors had masked repeat sequences in the GEO data set but that these sequences might remain in the Replication Domain data set. That the aforementioned emails confirmed that no additional smoothing or averaging or replication times had been performed by the

Replication Domain database, and made no mention of repeat masking, suggested that neither of these possibilities would explain the observed discrepancies.

Given that both data sets were allegedly based on assembly dm3 and, excepting 315 rows, should be identical, what then explained the discrepancy between the two data sets and which should have been used for the analysis? To attempt to answer this, chromosomal nomenclature was examined. The *D.melanogaster* genome contains heterochromatic regions. For naming purposes in assembly dm2 these regions were grouped together for each chromosome eg. chr2h. By assembly dm3, all of these regions had been assigned to individual chromosome arms and the naming convention updated to reflect this eg. chr2LHet and chr2RHet. Note also the change from 'h' to 'Het'. Examination of the two data sets showed that chromosomes in the GEO data set followed the former convention, suggesting that it was based on assembly dm2, whereas the Replication Domain data set used the latter chromosome names, indicating that probe positions were located on assembly dm3. Dr Schwaiger agreed that this was indeed the case, and that, contrary to previous communications and to the array based information supplied on the NCBI GEO database, the GEO data set was likely to have been based on assembly dm2.

The substitution rates used by Weber *et al*. (2012) in their main analyses were in part based on *D.melanogaster* exonic sequences extracted from file dmel-all-CDS-r5.33.fasta, available as a precomputed file from FlyBase. As these sequences were based on FlyBase release R5.33, it was therefore necessary to further convert the replication timing data to assembly 5, so that replication times could be correctly assigned to genes based on overlaps of genomic locations on a common assembly.

Due to the ambiguity associated with the initial source files, an attempt was made to regenerate the Replication Domain data set by converting the GEO data set from assembly dm2 to assembly dm3, to confirm that these data sets were indeed based on the suspected assemblies. The only conversion tool available to do this was the UCSC liftOver tool and associated chain file dm2ToDm3.over.chain. Lifting each probe position individually and a bulk lift with correct error handling both gave the same results. However, comparison of the lifted GEO data with the Replication

domain data revealed that 3,159,105 probe positions had been lifted but that only 306 probes generated errors, fewer than the 315 probes that Dr Hiratani had previously had to remove from the ReplicationDomain data set. It was possible that updates to the chain file since the data was originally lifted might explain why fewer errors were generated for the current lift. Further, only 3,127,385 probe positions were found to be common to each data set and of these, only 3,127,366 also had the same replication time, the remaining 19 common probes differing in their replication times. Whether these discrepancies might similarly stem from updates to the chain file or from another cause could not be determined.

Due to the differences identified between the lifted GEO data set and the Replication Domain data set, it was decided that complete re-curation of the original GEO data set was likely to yield more reliable results and that the re-lifted GEO data set should therefore be further converted to positions on assembly 5.

No UCSC liftOver chain file was yet available for lifting *D.melanogaster* positions to assemblies later than dm3, necessitating the use the FlyBase coordinate converter, which can convert positions between assemblies dm3, dm4 and 5 (Tweedie *et al*., 2009). Unlike the UCSC liftOver tool, no standalone command line driven version of the FlyBase coordinate converter was available so it was necessary to use the web based tool. As the website was not able to cope with the full data set, batches of 200,000 probe positions from assembly dm3 were reformatted, uploaded to the FlyBase coordinate converter and then their position in assembly 5 extracted from the resulting output file. However, a limitation of this process was that despite the FlyBase website providing assembly 5 positions for genes located in heterochromoatic regions, the coordinate converter unfortunately did not accept either heterochromatic or unknown chromosomal locations. As such, genes located in heterochromatic regions would not therefore be assigned replication times.

Examination of data lifted to assembly 5 revealed that of the 3,159,411 original probes, 148,178 failed to be converted to assembly 5. These were purged from the final data set. Of the 3,011,233 probes whose positions had been updated to assembly 5, 3,002,432 were found to occur uniquely at a given positions. The

remaining 8,801 probes were identified as duplicates, whereby a genomic location on assembly 5 had been assigned to more than one probe. At each of these positions, replication times were compared. In 7,258 cases, the replication time did not differ between the duplicate probes, in which case only a single probe (a total of 3,629 probes) was retained in the final data set.

In the remaining 1,543 cases, replication times were found to differ between probes assigned to a single position. Here, replication times for each pair were plotted. Identical replication times for two duplicate probes would fall along the line $y=x$. Where replication times differed, orthogonal residuals from $y=x$ would therefore provide a quantitative measure of the deviation from equality. These were calculated directly by application of Pythagoras' theorem to the right-angled isosceles triangle formed between the data point for the pair of probes and the line $y=x$. The orthogonal residual for each pair of probes was calculated as:

$$r = \frac{\sqrt{(y-x)^2 + (y-x)^2}}{2} \qquad (19)$$

where r is the orthogonal residual, x is the replication time of the first probe and y is the replication time of the second probe, shown graphically in Figure A1.1.

From the distribution of these residuals (Figure A1.2) an upper limit of 0.069 was imposed. For the 88 pairs or triplets of probes where the orthogonal residual exceeded 0.069 (Figure A1.3), both probes were purged from the final data set. In the 1,362 cases where orthogonal residuals were ≤ 0.069 (Figure A1.3), a mean of the two replication times was taken and assigned to a single probe at that position (Figure A1.3), thus retaining 681 probes. The final data set based on assembly 5 therefore contained 3,006,742 probes. The plots in Figure A1.4 compare the profile of these curated replication times along each chromosome in assembly 5 with the equivalent replication timing profiles based on assembly dm3, published on the Replication Domain database.

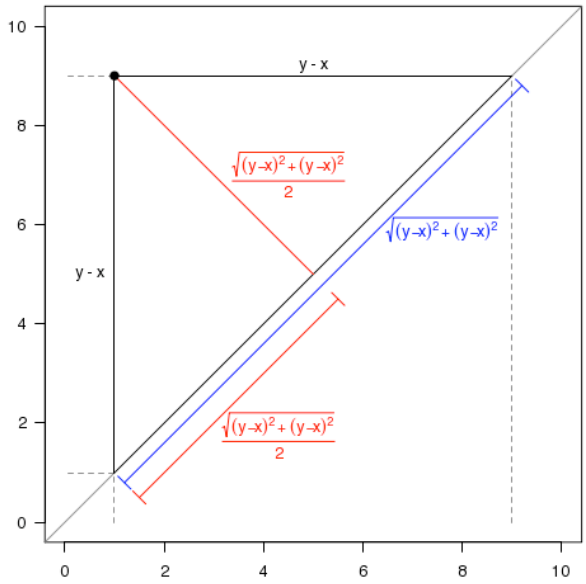


Figure A1.1: Graphical representation of the method used to calculate orthogonal residuals of points from the line $y=x$. In this simplified case, the replication times of the two probes are 1 and 9. The point (x=1 and y=9) sits at the right-angled corner of an isosceles triangle, with two sides both equal to $y–x$. The orthogonal residual, shown in red, is equal to half of the hypotenuse, shown in blue.

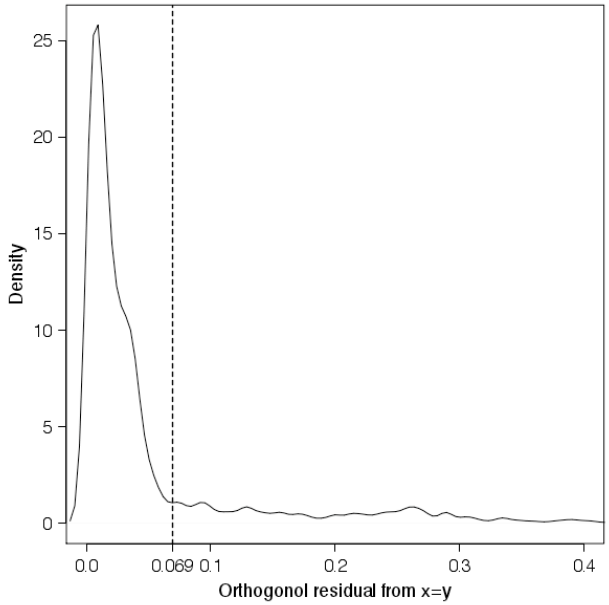


Figure A1.2: Distribution of orthogonal residuals of duplicate replication times from $y=x$. Dashed line at 0.069 represents the maximum permitted residual used to filter duplicate replication times.

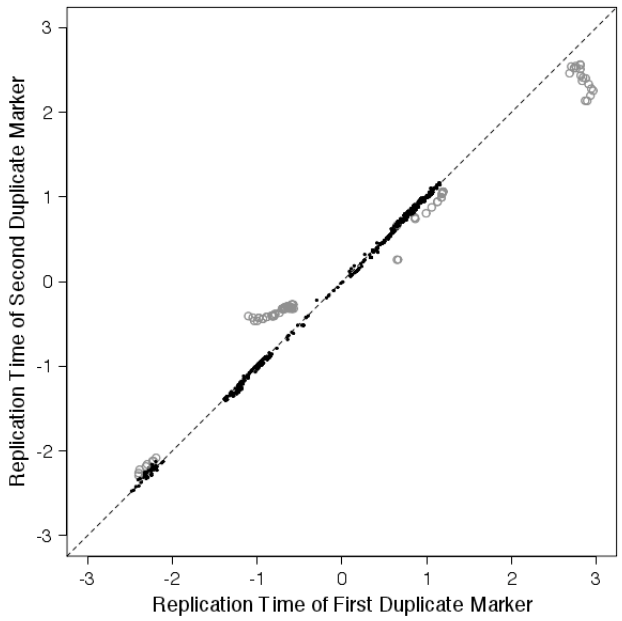


Figure A1.3: Plot of different replication times assigned to single probe positions. Dashed line is $y=x$. Closed black dots represent probes with orthogonal residuals ≤ 0.069 for which a mean replication time was taken and assigned to the position. Open grey circles represent probes with orthogonal residuals > 0.069, in which cases both probes were purged from the final data set.

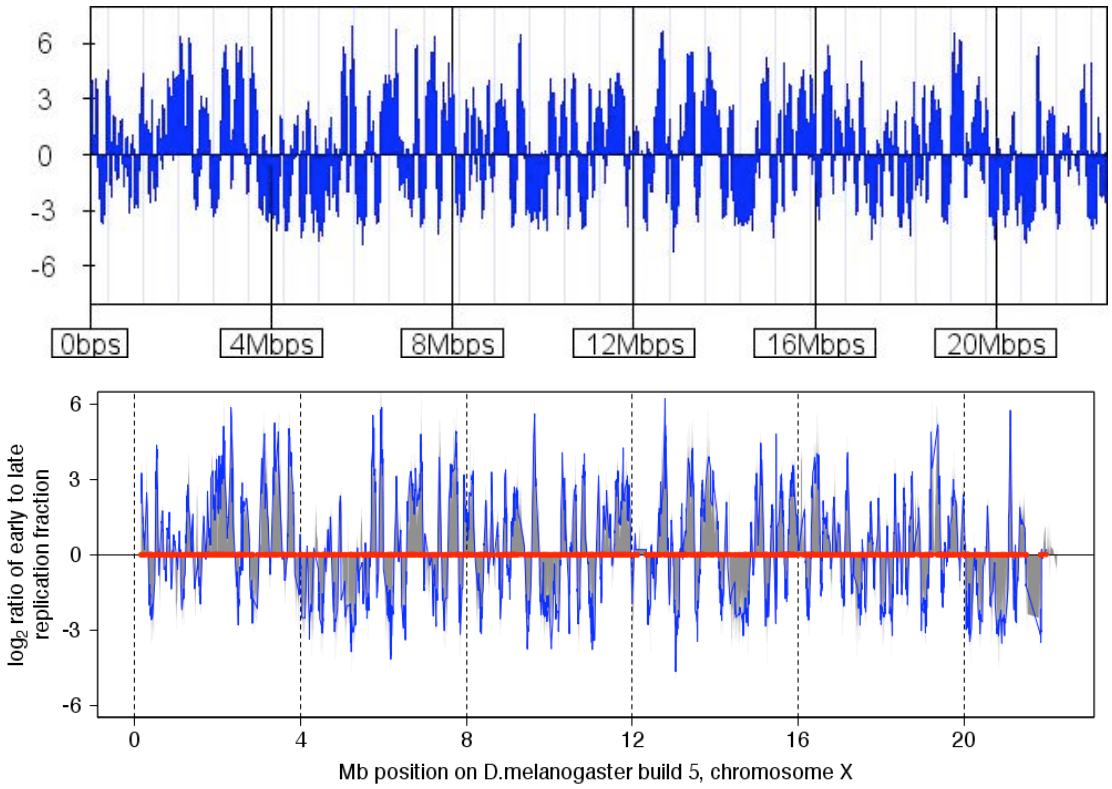
Finally, replication times were assigned to genes based on identification of all probes located within a given gene. Where more than one replication time was identified for a given gene, these were tested for normality of distribution, 23% of which were found to be skewed. As such, the median of all replication times that applied to a given gene was taken. The distributions of these 22,689 genes, together with their assigned genic replication times (Figure A1.4) suggest that genes are sampled from both early and late replicating domains.

To conclude, curation of this data set of replication timing in *D.melanogaster* demonstrates why careful consideration of the assembly on which data sets are generated and the nature of data obtained from external sources is required. The increasing availability of genome-wide data for an expanding range of genome features, coupled with updated genome assemblies, particularly for newly sequenced species, mean that these types of issues are likely to be encountered more frequently in the future. Findings obtained from analyses of a range of genome features should therefore be treated with caution if such issues have not been addressed in the

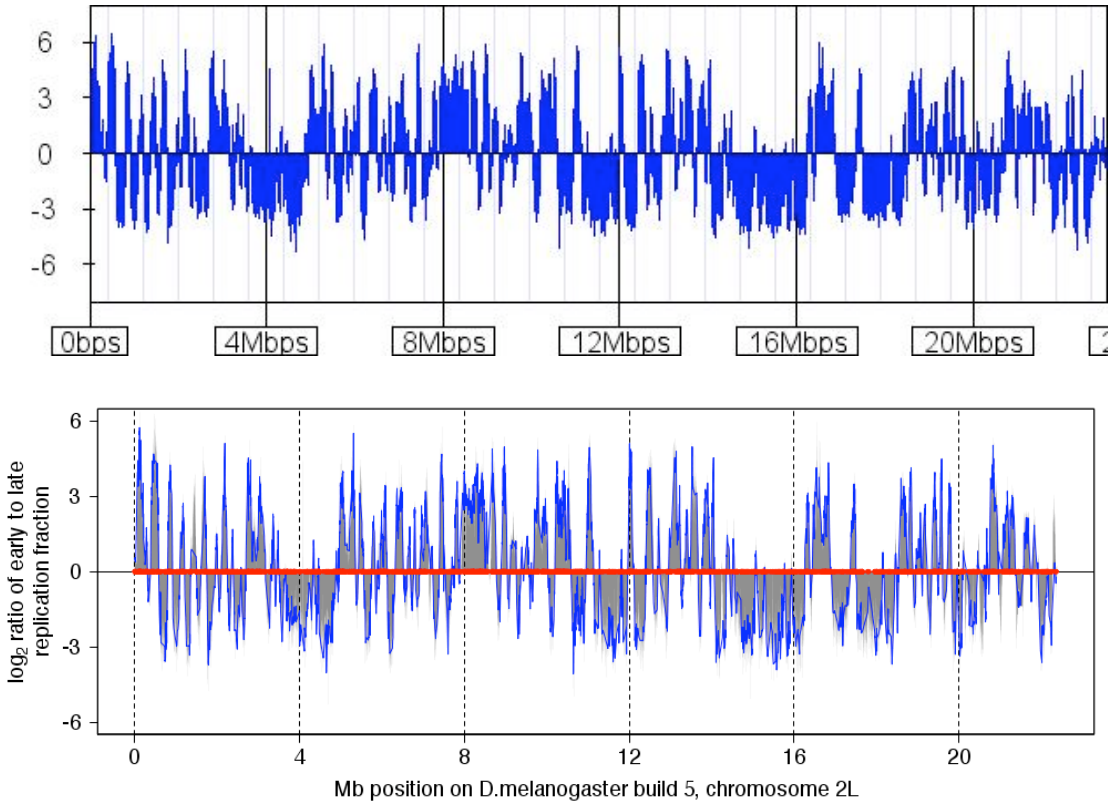
methodologies. It is interesting to note that the UCSC repository has recently added additional flags to its genome browser indicating which tracks have been lifted from previous assemblies.

Figure A1.4: Distribution of replication times along *D.melanogaster* chromosomes. For each pair of plots, the top image is taken from Replication Domain and represents timings based on assembly dm3. The bottom image is generated from the curated data set based on assembly 5. The grey shaded region shows probe replication times. Red dots indicate the position of genes along the chromosome. The blue line shows the replication time assigned to these genes.

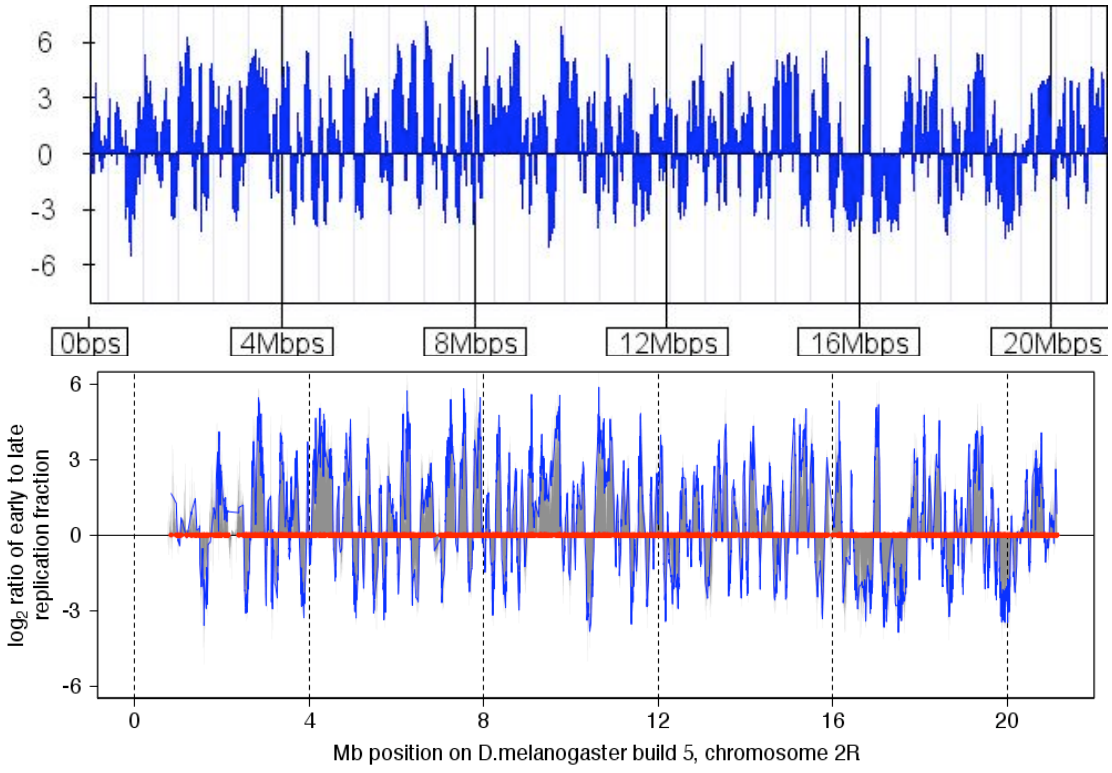
Chromosome X:



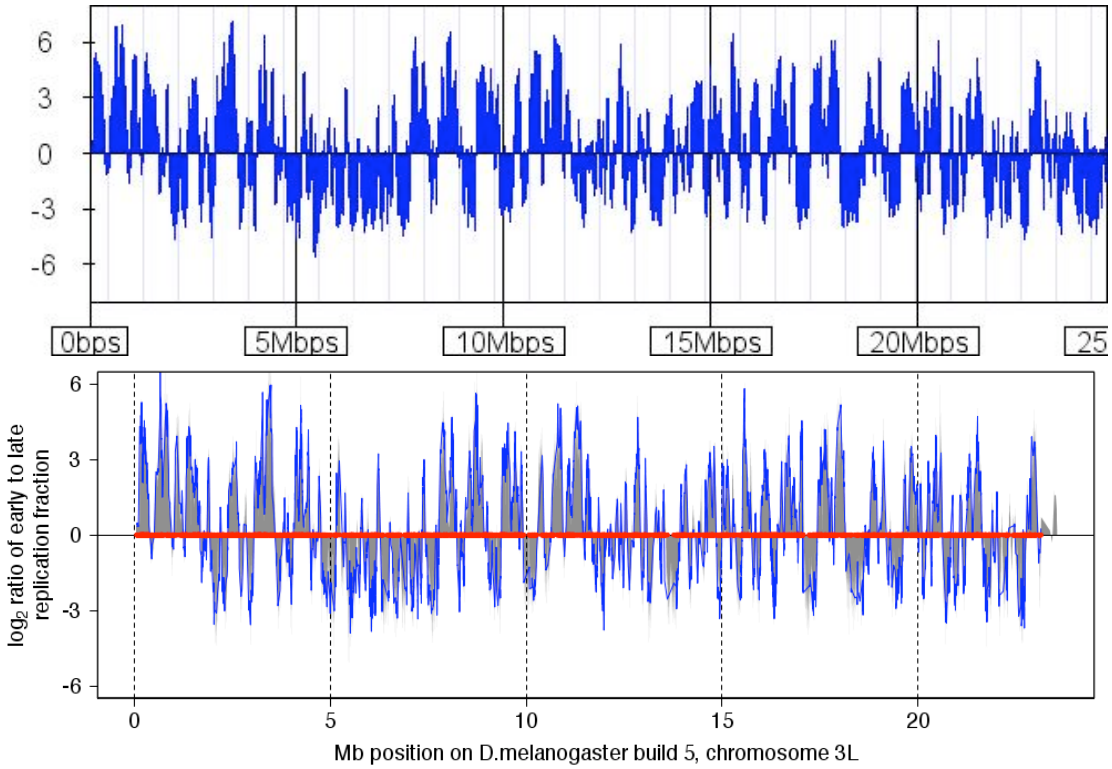
Chromosome 2L:



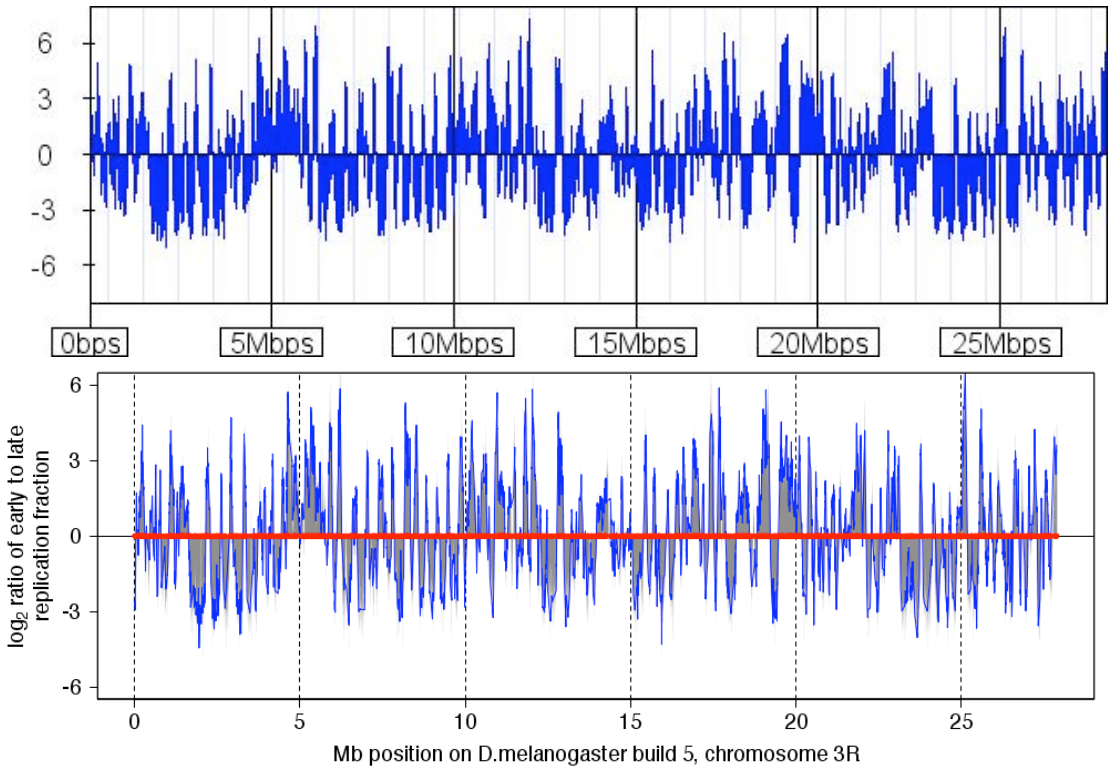
Chromosome 2R:



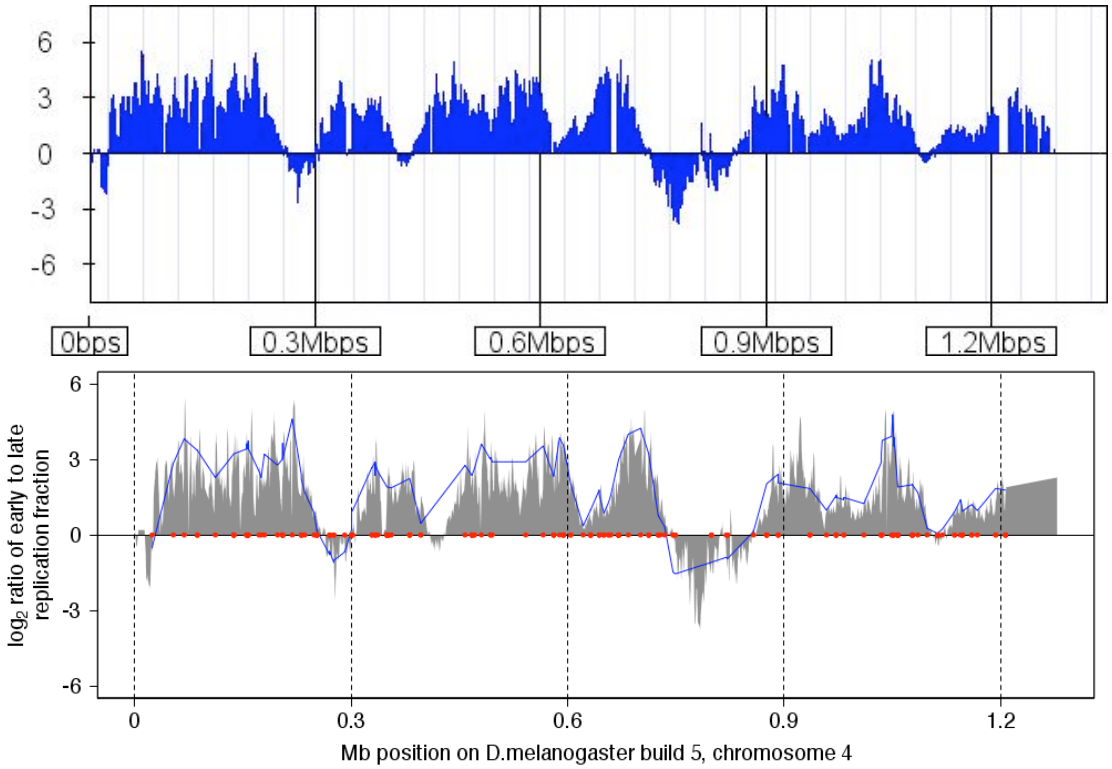
Chromosome 3L:



Chromosome 3R:



Chromosome 4:



A1.2 References

CHEN, C.-L., RAPPAILLES, A., DUQUENNE, L., HUVET, M., GUILBAUD, G., FARINELLI, L., AUDIT, B., D'AUBENTON-CARAFA, Y., ARNEODO, A., HYRIEN, O. & THERMES, C. (2010) Impact of replication timing on non-CpG and CpG substitution rates in mammalian genomes. *Genome Research,* 20**,** 447-457.

GILBERT, D. M. (2010) Evaluating genome-scale approaches to eukaryotic DNA replication. *Nat Rev Genet,* 11**,** 673-684.

MACALPINE, D. M., RODRÍGUEZ, H. K. & BELL, S. P. (2004) Coordination of replication and transcription along a *Drosophila* chromosome. *Genes Dev,* 18**,** 3094-3105.

PINK, C. J. & HURST, L. D. (2010) Timing of replication is a determinant of neutral substitution rates but does not explain slow Y chromosome evolution in rodents. *Mol Biol Evol,* 27**,** 1077-1086.

SCHWAIGER, M., STADLER, M. B., BELL, O., KOHLER, H., OAKELEY, E. J. & SCHÜBELER, D. (2009) Chromatin state marks cell-type- and gender-specific replication of the *Drosophila* genome. *Genes Dev,* 23**,** 589-601.

STAMATOYANNOPOULOS, J. A., ADZHUBEI, I., THURMAN, R. E., KRYUKOV, G. V., MIRKIN, S. M. & SUNYAEV, S. R. (2009) Human mutation rate associated with DNA replication timing. *Nat Genet,* 41**,** 393-395.

TWEEDIE, S., ASHBURNER, M., FALLS, K., LEYLAND, P., MCQUILTON, P., MARYGOLD, S., MILLBURN, G., OSUMI-SUTHERLAND, D., SCHROEDER, A., SEAL, R., ZHANG, H. & CONSORTIUM, F. (2009) FlyBase: enhancing *Drosophila* Gene Ontology annotations. *Nucleic Acids Res,* 37**,** D555-D559.

WEBER, C. C., PINK, C. J. & HURST, L. D. (2012) Late-Replicating Domains Have Higher Divergence and Diversity in *Drosophila melanogaster*. *Molecular Biology and Evolution*. 29, 873–882.

Appendix 2: **Published Papers**

Evidence that replication-associated mutation alone does not explain between-chromosome differences in substitution rates

Catherine J. Pink, Siva K. Swaminathan, Ian Dunham, Jane Rogers, Andrew Ward, and Laurence D. Hurst

Genome Biology and Evolution (2009). 1(1): 13-22

Timing of replication is a determinant of neutral substitution rates but does not explain slow Y chromosome evolution in rodents

Catherine J. Pink and Laurence D. Hurst

Molecular Biology and Evolution (2010). 27(5): 1077-1086

Late replicating domains are highly recombining in females but have low male recombination rates: Implications for isochore evolution

Catherine J. Pink and Laurence D. Hurst

PLoS One (2011). 6(9): e24480